Algorithms, Biases, and Belief Polarization*

Varad Deolankar† Jessica Fong‡ S. Sriram§

February 28, 2026

Preliminary draft. Please do not cite or circulate.

Abstract

The rise of personalized engagement-optimizing recommendation systems has raised concerns about their role in driving polarization, but disentangling their impact from intrinsic human tendencies remains challenging. This paper investigates two pathways to polarization: (a) information gatekeeping by algorithms that curate content to maximize engagement and (b) biases in information processing. We further investigate how different engagement metrics, such as dwell time and likes, shape content exposure and escalate polarization. Our four-step approach combines (1) an online experiment tracking participants' belief updating and engagement decisions, (2) structural model estimation to capture information processing biases, (3) deep reinforcement learning agents trained using the estimated behavioral model to emulate personalized recommendation systems, and (4) an evaluation of the belief structures resulting from the interaction between the trained recommendation agents and the estimated structural model. Our method allows us to infer the relative impact of personalized engagement-optimizing recommendation algorithms vs. biases in information processing on belief polarization. Our findings suggest that information processing biases and algorithmic curation each amplify polarization to a comparable degree. Among information processing biases, weight distortion—the tendency to discount belief-incongruent information—is the primary channel through which information processing biases escalate polarization, while signal distortion—the tendency to misinterpret belief-incongruent information—has a negligible net effect. Both like-maximizing and dwell-time-maximizing algorithms increase polarization relative to random content assignment, but like-maximization generates more polarization by serving belief-confirming content that anchors users with extreme priors. Dwell-time maximization, by contrast, more frequently exposes moderate users to content from the opposite side of the attitudinal spectrum.

*We are thankful for comments from Ali Goli, Pradeep K. Chintagunta, Jun Li, Zach Y. Brown, Katherine Burson, Scott Rick, Hulya Karaman, Nanda Kumar, and Madhu Vishwanathan. All errors are our own. The usual disclaimer applies.

†National University of Singapore, varad.deolankar@nus.edu.sg

‡University of Michigan, jyfong@umich.edu

§University of Michigan, ssrira@umich.edu

Keywords— Recommendation Systems, User Engagement, Polarization, Platform Design, Information Processing Biases

1 Introduction

There has been growing concern about societal polarization across the world (Pew Research Center, 2014; Carother and O'Donohue, 2019; Murray and Marquez, 2022). This trend is usually attributed to the contemporaneous rise in digital media consumption (Törnberg, 2022; Arora et al., 2022). There are two potential pathways through which polarization could occur: (a) what information is made available and (b) how information is processed (Dahlgren, 2020; Modgil et al., 2024). The first pathway is closely tied to the growth in digital media consumption. A widely discussed explanation is that engagement-driven recommendation algorithms tend to prioritize content that aligns with users' existing views. By systematically curating information environments in ways that reduce exposure to opposing perspectives, such algorithms may reinforce prior beliefs. Over time, this selective exposure can contribute to increasingly entrenched and divergent positions across groups.

The second pathway relates to how individuals process information and update their beliefs. A growing body of research documents systematic biases in information processing (Freedman and Sears, 1965; Frey, 1986). Individuals often reinterpret conflicting evidence to fit existing beliefs (Rabin and Schrag, 1999) and assign lower credibility to belief-incongruent information (Mehta et al., 2008). We refer to these tendencies as *signal distortion*—the biased perception of the informational content itself, whereby the interpretation of a signal systematically shifts depending on its congruence or incongruence with one's prior beliefs—and *weight distortion*—the differential weighting of information based on its alignment with prior beliefs. In the presence of these biases, polarization can arise even absent algorithmic curation. This pathway provides an alternative explanation for polarization that is not directly tied to the growth in digital media consumption and the algorithms these platforms employ.

In this paper, we investigate the drivers of belief polarization by examining two distinct yet potentially complementary mechanisms: biases in how individuals process information and update their beliefs, and the composition of information made available through personalized recommendation algorithms. By jointly modeling individual-level information-processing distortions and algorithmic curation, we assess their relative contributions to the amplification of belief differences. Second, we examine how algorithmic amplification of polarization depends on the engagement metric used to optimize recommendation systems. In practice, platforms train algorithms to maximize specific behavioral signals, such as explicit feedback (e.g., "likes") or implicit measures (e.g., dwell time).[1] The

[1]For example, TikTok likely maximizes dwell time (https://www.reuters.com/technology/, accessed

choice of objective function may shape the informational environment users encounter, as content that is "liked" may differ from content that sustains attention. We therefore analyze whether alternative engagement metrics generate differential patterns of exposure and, in turn, have differential effects on polarization.

To achieve these research objectives, we implement a four-step methodology. First, we run an online experiment that simulates a content browsing session. Participants are sequentially exposed to short articles about the harmfulness of genetically modified organisms (GMOs), a topic characterized by substantial heterogeneity in prior beliefs (Fong et al., 2024). In each period, the displayed article is randomly drawn from a set of GPT-4–generated articles that vary in their direction (pro vs anti-GMOs) and extremity. For each exposure, we collect three key measures: (i) participants' interpretation of the article's stance (i.e., the extent to which they perceive it as supporting GMOs), (ii) their beliefs about GMO harmfulness before and after reading the article, and (iii) their engagement behavior, namely, dwell time and reactions (like, dislike, or no response). These data allow us to observe how individuals interpret information, update beliefs, and engage with content conditional on their prior beliefs.

Second, we use these data to estimate a model that characterizes how users adjust their beliefs based on new information and engage with the content. This model augments Bayesian updating with the two channels through which biases in information processing can occur—signal and weight distortion. In the third stage, we use the estimates from this structural model to generate synthetic data to train deep reinforcement learning agents used in personalized engagement-optimizing recommendation systems.[2] Here, we leverage our estimated model to mimic the environment, providing a behaviorally grounded foundation for training recommendation agents. Finally, we deploy the trained recommendation agents and simulate belief dynamics under different scenarios—varying the recommendation algorithm (random, like-maximizing, or dwell-time-maximizing) and selectively enabling or disabling information processing biases. By comparing posterior beliefs across these scenarios, we can disentangle the contribution of algorithmic content curation from that of information processing biases in driving belief polarization.

We find that individuals exhibit both signal distortion and weight distortion. The magnitude of the distortions vary based on whether the information is consistent with an individual's prior belief

February 2026).

[2]Prior research on recommendation systems has used trained models to mimic the environment and train agents (Ha and Schmidhuber, 2018; Chen et al., 2021).

and more extreme vs. moderate, or in the opposite direction. Our results suggest that individuals distort confirmatory content (signals in the same direction as the prior) towards their prior beliefs; we do not find that individuals systematically misinterpret disconfirmatory signals as confirmatory. Weight distortion varies based on the position of the content relative to the prior as well: the lowest weight distortion (i.e., placing less weight in new information relative to the Bayesian benchmark) occurs for content that is closely aligned with an individual's prior belief. As a result, individuals place less weight on more extreme confirmatory signals and disconfirmatory signals. Thus, weight distortion dampens shifts in beliefs.

The alignment of content with an individual's prior beliefs affects their engagement with the content, though different engagement metrics respond in distinct ways. Individuals are more likely to "like" attitudinally-proximal articles and to "dislike" attitudinally-distant ones. In contrast, they spend more time, on average, on articles that are closer to the center on either side of the pro- and anti-GMO spectrum. As a result, recommendation algorithms optimized for different engagement objectives will prioritize different types of content. A dwell-time-maximizing algorithm will recommend mildly opposing or more moderate articles, while a like-maximizing algorithm will recommend content close to the individual's prior.

We explore the relative roles of information-processing biases and algorithmic curation in driving belief polarization using a series of counterfactual simulations. We model belief dynamics over a ten-period browsing session under alternative recommendation regimes—random assignment, like-maximization, and dwell-time–maximization—and compare outcomes with and without information-processing biases. We find that both information processing biases and algorithms independently increase polarization. Among the processing biases, weight distortion is the primary driver: by systematically discounting disconfirming information, it limits updating toward more moderate beliefs. In contrast, signal distortion has a negligible net effect because it symmetrically shifts perceived signals toward prior beliefs, generating offsetting polarizing and moderating forces.

We then quantify the role of the algorithm across various engagement objectives in belief polarization. Both like-maximizing and dwell-time-maximizing algorithms increase polarization relative to random content assignment. However, like-maximization generates substantially more polarization than dwell-time maximization. This difference arises because the two algorithms serve fundamentally different content. Like-maximization delivers belief-confirming content to users with extreme priors, anchoring them in their existing positions. Dwell-time maximization, by contrast, more frequently exposes moderate users to content from the opposite side of the attitudinal spectrum. Overall,

the magnitudes of algorithmic and psychological drivers are comparable. Information-processing biases increase polarization by 17.76%, while algorithmic optimization increases polarization by 9.73% under dwell-time maximization and 17.04% under like-maximization.[3] These findings suggest that polarization cannot be attributed solely to algorithmic curation, as individual-level processing biases play an equally consequential role. Furthermore, we find that the higher polarization due to biases is mostly concentrated among individuals starting with extreme beliefs, while the algorithms result in higher polarization among individuals who start with moderate beliefs. Thus, biases and algorithms lead to higher polarization through different pathways.

These findings have implications for policymakers, firms, and consumers. For instance, we find that feeds curated by algorithms that attempt to maximize dwell time polarize beliefs to a lower extent than those maximizing likes. Note that engagement measures such as dwell time rely on cookies, which are being increasingly regulated due to privacy concerns, whereas measures such as "likes" do not rely on cookies. We find that a more invasive engagement measure such as dwell time leads to lower belief polarization compared to a less invasive measure such as "likes." This implies that allowing firms to collect dwell time data and use it in their recommendation engines could be desirable. Hence, understanding the impact of the engagement metric that the platform chooses to optimize on the polarization of beliefs can provide guidance for shaping privacy protection policies.

The rest of the paper is organized as follows. Section 2 discusses the related literature. Section 3 describes our experimental design and data collection. Section 4.1 presents the structural model of information processing and content engagement. Section 5 reports the model estimates and describes the simulation method. Section 6 decomposes the drivers of polarization through counterfactual simulations. Section 7 concludes with the implications of our findings, the limitations of our study, and fruitful avenues for future research.

2 Related Literature

Belief polarization — the tendency for individuals' opinions to become more extreme and more divided over time — has been studied across psychology (Itzchakov and Van Harreveld, 2018), political science (Kim and Kim, 2019; Yarchi et al., 2021), and information systems (Modgil et al., 2024). The literature identifies two primary mechanisms through which polarization can arise:

[3]Each algorithm was calibrated to generate a 5% increase in its respective engagement metric relative to random assignment, implying that more aggressive optimization could yield even larger polarization effects.

exposure effects, driven by the information environment individuals encounter, and processing effects, driven by the cognitive biases individuals bring to that information.

A large body of work exists on exposure effects. This literature typically argues that platform recommendation algorithms narrow users' information environments in ways that reinforce pre-existing beliefs. Sunstein (2009) and Pariser (2011) articulate the theoretical case for "filter bubbles," in which personalized feeds systematically reduce exposure to cross-cutting viewpoints. Consistent with this view, Levy (2021) document that social media algorithms limit exposure to counter-attitudinal news and thereby increase polarization, and Huszár et al. (2022) find evidence of systematic amplification of politically aligned content across multiple platforms.

Other work, however, finds weaker or null effects of algorithmic curation. Bakshy et al. (2015) show that individual choice — not the algorithm — is the primary driver of ideologically homogeneous news exposure on Facebook. Guess et al. (2023) conduct a large-scale randomized experiment during the 2020 U.S. election, finding that switching Facebook feeds from algorithmic to chronological order did not meaningfully change political attitudes or polarization, despite substantially altering content exposure. Dandekar et al. (2013) offer a theoretical result showing that PageRank-style relevance algorithms are always polarizing in expectation, while Cinelli et al. (2021) demonstrate that the extent of echo chamber formation varies substantially by platform architecture. We contribute to the literature on algorithmic drivers of polarization by demonstrating that the specific engagement metric that is optimized has meaningful consequences for belief polarization.

A parallel literature examines how individuals process information independently of what they are exposed to. The foundational work on confirmation bias — the tendency to seek out, interpret, and recall information in ways consistent with prior beliefs — is reviewed in Freedman and Sears (1965) and Frey (1986). Within this broad tendency, researchers distinguish between two specific mechanisms that are directly relevant to belief updating. First, *signal distortion* refers to the systematic misinterpretation of ambiguous or opposing evidence as consistent with one's priors (Rabin and Schrag, 1999). Second, *weight distortion* refers to the selective discounting of belief-incongruent information at the updating stage (Mehta et al., 2008; Fong et al., 2024). This paper contributes by separately identifying and estimating each channel, by showing that weight distortion — not signal distortion — is the primary bias mechanism driving belief polarization, and by comparing the role of these biases to that of the algorithm in driving polarization.

3 Method

In order to study how personalized recommendation algorithms impact beliefs, we need to understand the following: (1) how individuals update their beliefs in response to new information, (2) how the information and their beliefs impact their engagement decisions, and (3) how information and the order in which it is presented influences the distribution of beliefs in the population. Within (1), to account for the potential role of biases in information processing—where individuals may favor information reinforcing their preexisting beliefs—it is also necessary to examine how they perceive the information and how these perceptions diverge from the objective content of the information itself. Such information is not usually available in observational data. Therefore, we perform an experiment, wherein we collect data that would help us infer (1) and (2). Furthermore, to assess the role of information-processing biases, one would ideally compare belief polarization in a real-world scenario to a counterfactual scenario in which such biases are absent. However, it is difficult to construct this counterfactual experimentally, as we cannot directly turn off individuals' inherent tendencies in interpreting information and updating beliefs. Therefore, we calibrate a structural model to generate counterfactuals that would help us with (3). In light of these considerations, we propose a four-step methodology as follows:

- **Step 1:** Run an experiment where we present content randomly to users and collect data on their belief adjustment and engagement decisions (the decision of how much time to spend and the decision to like, dislike, or not react).
- **Step 2:** Use these data to estimate a structural model of users' belief adjustment and engagement decisions.
- **Step 3:** Use the calibrated structural model generate synthetic data to train deep reinforcement learning agents that mimic personalized engagement-optimizing recommendation systems.
- **Step 4:** Assess the differences in belief structures resulting from the interaction between trained recommendation agents and the estimated structural model for different counterfactual scenarios of interest.

In the sections below, we describe each of these steps in detail.

3.1 Experiment Design

In the experiment, we collect data that allows us to observe how individuals update their beliefs in response to new information, and how the information and their beliefs impact their engagement decisions. We do so by having participants engage in a content browsing session, where they are exposed to several short articles about the harmfulness of Genetically Modified Organisms (GMOs). We choose GMOs because it is a controversial topic with a variety of prior beliefs about its harmfulness (Fong et al., 2024). We randomly select the articles and the order they are presented from a repository of 200 articles that we created using GPT-4. Half of these articles argue that GMOs are not harmful, while the other half argue that GMOs are harmful. Within each sentiment group, the articles also vary in intensity. We specify an intensity score, which ranges from 0.01 to 1.00, in the prompt provided to GPT-4 while generating the article.[4]

Prior to the content browsing session, we elicit prior beliefs about the harmfulness of GMOs as well as the uncertainty our participants have in their prior belief as shown in Figure A.1. After viewing each article, we ask participants about two key beliefs: (a) their belief of the recommended article's stance towards GMOs, and (b) their belief about the harmfulness of GMOs after reading the article. By examining both the true signal and the perceived signal as well as the prior and posterior beliefs, we can quantify the extent to which individuals incorporate their perception of new information when updating their beliefs. Specifically, we investigate whether a systematic relationship exists between the degree of discrepancy between new information and a participant's prior belief and (i) the deviation of the perceived signal from the truth (i.e., signal distortion), and (ii) the weights individuals assign to the perceived signal (i.e., weight distortion). This allows us to assess the role of both signal and weight distortion.

Additionally, participants have the option to engage with the article by liking or disliking it, as they would on a digital media platform, or choose not to engage at all. We also observe the time each participant spends reading the article. This allows us to observe how beliefs and content impact participants' engagement decisions. Figure A.2 shows an example of an article and how we solicit beliefs and engagement.

[4]We asked GPT-4 to create an article with a given intensity of X using the following prompt: "Can you generate a one-sentence long headline followed by a paragraph-long piece of text that argues in favor of the use of Genetically Modified Organisms (GMOs) in common food products with an intensity level of X. Please assume that intensity is a metric on a scale from 0 to 1 where 0 is not intense and 1 is extremely intense. Please be sure to not mention anything about the intensity score in the generated text." Similarly, we also created articles that argue against the use of GMOs.

A challenge in soliciting self-reported beliefs is that participants may not always tell the truth. For example, on socially sensitive issues, individuals might report answers that they expect the researcher will view as desirable rather than their true answers (Paulhus, 1991). In such settings, researchers have designed questions to be incentive-compatible with reporting the truth by rewarding participants whose answers are correct. However, this is not feasible in our setting because we are interested in participants' beliefs, not what they think is the correct answer. Nevertheless, to protect against biases stemming from the use of self-reported single-item measures, we use the error choice technique to elicit beliefs (Hammond, 1948; Ajzen et al., 2018) and collect data in two waves. In the first wave, we collect data on participant demographics. In the second wave, administered the next day, we expose the participants to the content browsing session previously described. We ask participants to put themselves in the shoes of another participant with the same demographic profile and answer our questions from the perspective of this "other participant." The idea behind this approach is that the participant will provide answers that are consistent with their own beliefs because they have similar demographic traits as the "other participant" (Krosnick et al., 2005).

3.2 Data Description

We recruited 1500 U.S. participants from Prolific, using their representative sampling tools to obtain a sample reflective of the U.S. population.[5] In Appendix A.2, we report the demographics of the participants in our study. We collected these data in the first wave of our study and used them for the second wave.

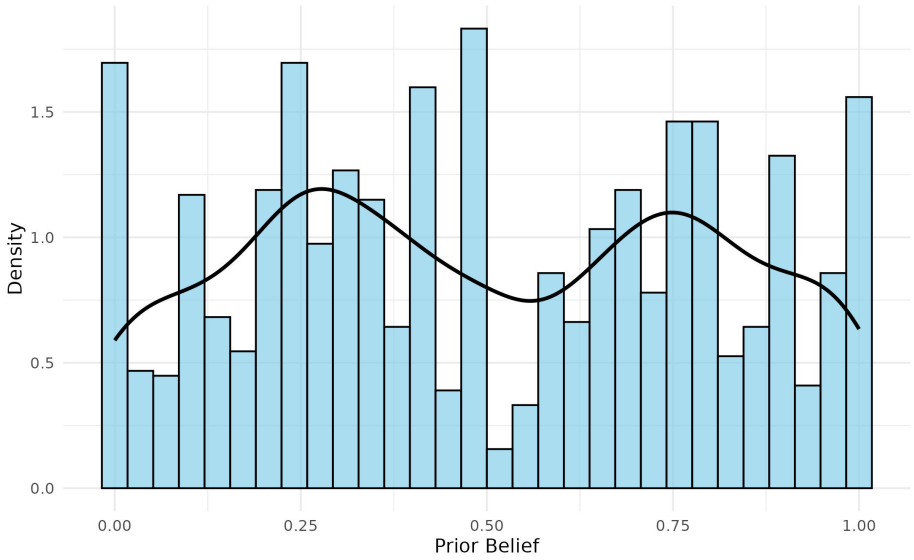
Figure 1 plots the distribution of the strength of prior beliefs about the harmfulness of GMOs. As we expected in the context of GMOs, there is variation in prior beliefs. Figure 2 plots the uncertainty of the prior beliefs. The Figure reveals that there is also considerable variation among respondents in terms of the confidence in their beliefs and that the degree of uncertainty is similar for the pro- and anti-GMO groups.[6] This variation in both the strength and uncertainty of prior beliefs enables us to examine heterogeneous effects by priors.

Table 1 describes the data on how participants interpret the signal, how they update their beliefs, and their engagement decisions. The first row shows that the average deviation between

[5]The participants were paid a fixed amount of $3.00 after they completed both waves of our survey. 1488 out of 1500 participants recruited in the first wave completed both waves of our survey experiment.

[6]An asymptotic two-sample Kolmogorov–Smirnov test indicates no statistically significant difference in the distributions of prior variance between "Anti-GMO" and "Pro-GMO" believers ($D = 0.0607$, $p = 0.1289$).

Figure 1: Distribution of Prior Beliefs



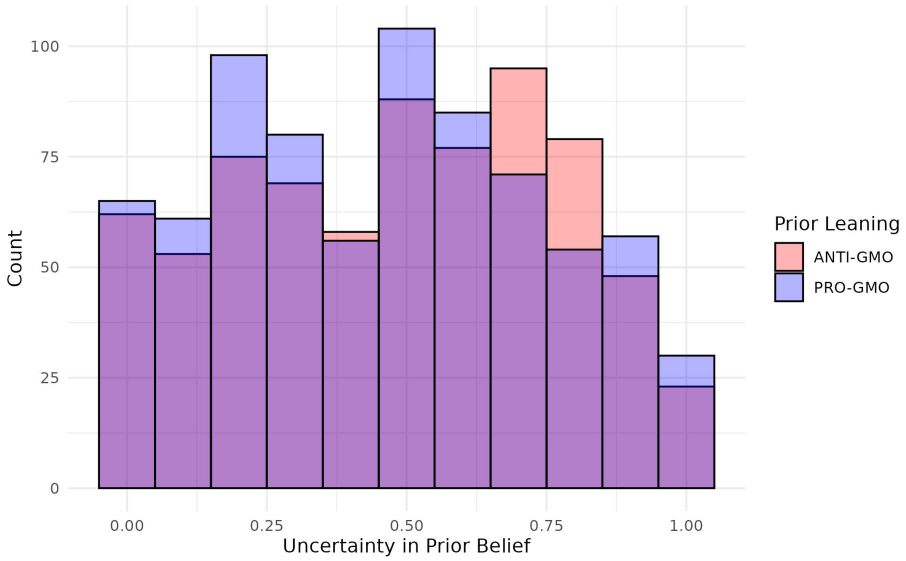
Notes. Prior Belief is measured using the first slider scale shown in Figure A.1. On a scale from 0 to 1, we measure how strongly the individuals believe in the hypothesis: "GMOs are not harmful". Thus, values closer to 0 suggest that individuals have a more "Anti-GMO" prior belief, whereas, values closer to 1 suggest that individuals have a more "Pro-GMO" prior belief. The black line overlays a kernel density estimate, which provides a smoothed approximation of the distribution.

the perceived and true signals is not zero.[7] The non-zero difference between the true and perceived signal indicates individuals interpret the true signal with some error. The second row shows that the recommended information shifts participants' beliefs, with the mean shift in beliefs, as calculated as the absolute difference between the prior belief before the session and posterior belief after the session, being 0.34.

As noted above, our data suggest that individuals interpret the pro- vs. anti-GMO slant of the articles with some error. In order to understand the source of this error, we bin the true signal into ten categories and plot the average perceived signal for each bin in Figure 3. Note that true signals that are greater than 0.5 indicate pro-GMO articles, while those than 0.5 are anti- GMO. Figure 3 shows two key patterns. First, we observe a clear monotonic increase, indicating that individuals are progressively more likely to perceive articles as pro-GMO as the underlying content becomes more pro-GMO. As a result, we find that the correlation between the perceived signal and the true signal

[7]Note that we observe both the perceived signal and the true signal. We collect the perceived signal in every period as shown in Figure A.2, while the true signal is the intensity score specified in the prompt provided to GPT-4 while generating the article.

Figure 2: Distribution of Uncertainty in Prior Beliefs



Notes. Uncertainty about prior belief is measured using the second slider scale shown in Figure A.1. On a scale from 0 to 1, we measure how uncertain individuals are in their belief about the harmfulness of GMOs. Values closer to 0 mean that individuals are less uncertain, whereas, values closer to 1 suggest that individuals are more uncertain.

Table 1: Summary statistics

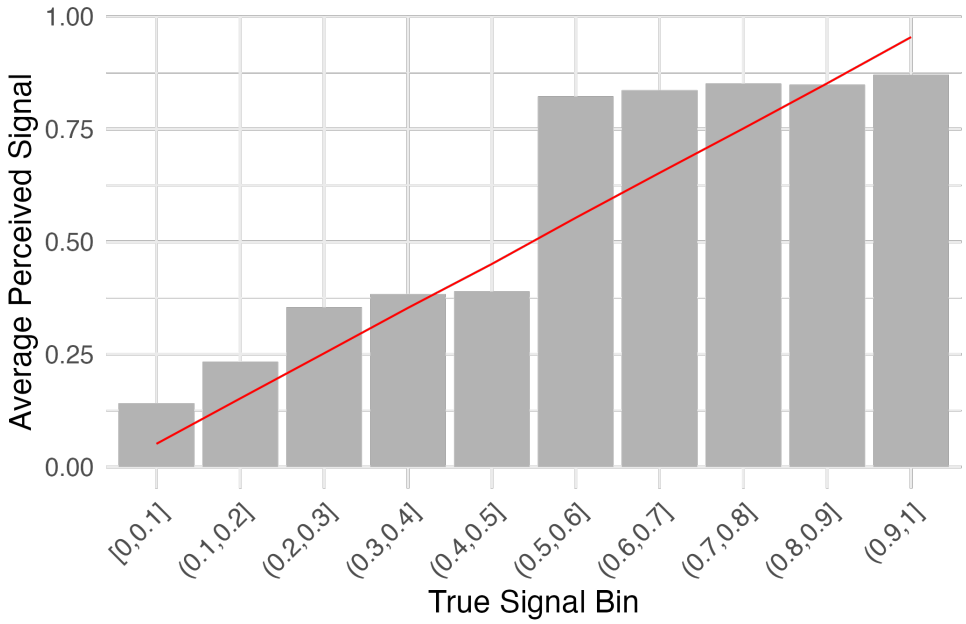
Statistic	**Mean**	**SD**	**Median**	**Min**	**Max**
Avg. distance between perceived signal and true signal	0.177	0.067	0.164	0.040	0.585
Belief shift between start and end of the browsing session	0.339	0.233	0.310	0.000	1.000
Avg. proportion of "Likes"	0.556	0.334	0.600	0.000	1.000
Avg. proportion of "NoReacts"	0.088	0.226	0.000	0.000	1.000
Avg. proportion of "Dislikes"	0.356	0.330	0.300	0.000	1.000
Avg. time spent reading the article (seconds)	38.305	28.878	31.588	1.127	304.325

Notes. For each individual, the average distance between perceived signal and true signal is computed as the average of the absolute value of the difference between the perceived signal and the true signal within the browsing session. Belief shift is the absolute value of the difference between the prior belief before the first period and the posterior belief after the last period in the browsing session. Average engagement statistics are also computed as the average engagement within the browsing session.

is high at 0.77. Second, the figure shows individuals tend to perceive content as more pro-GMO that it actually is, as average perceived signal is usually higher than the average true signal (represented by the line). The misperception is greater for pro-GMO content; individuals perceive this content to be more extreme than it is. In our empirical analysis, we conceptualize signal distortion as a systematic bias whereby perceived information deviates from the true signal as a function of the

distance between the true signal and the individual's prior belief. In order to isolate this pattern, we will include signal bin fixed effects, which will capture the average misperception for each bin.

Figure 3: Average perceived signal by true signal



Notes. The x-axis partitions the true signal $s_{it}^T \in [0,1]$ into equally spaced bins (deciles). Bars show the mean perceived signal s_{it}^P within each true-signal bin. The red line overlays the mean true signal within each bin. The vertical gap between the bar height and the red line in a bin represents the average misperception $(s_{it}^P - s_{it}^T)$ at that true-signal level.

The final four rows of Table 1 describe users' engagement with the recommended content. On average, participants "like" 55.6% of the recommended articles, do not react 8.8% of the time, and "dislike" 35.6% of the articles. At first glance, these content engagement rates may seem high. However, existing literature has shown that polarizing content generally receive higher engagement rates than neutral content (Kim and Ihm, 2020; Rathje et al., 2021, 2024), which might explain the high rates in our context. Lastly, participants spent 38.3 seconds, on average, reading each article. Participants spent more time reading articles with more words—a 1% increase in the number of words is associated with a 0.72% increase in reading time (Table B.1), suggesting that participants actually read the articles.

Table 2 presents summary statistics for dwell time and engagement behavior by whether participants perceived the article as aligned or misaligned with their prior beliefs. We define an article as perceived aligned if the participant's interpretation of the article's stance is on the same side of the

Table 2: Engagement by Alignment with Prior

Variable	Perceived Alignment	N	Mean	Median	SD
Dwell Time (sec)	Aligned	6874	38.29	25.73	64.38
	Misaligned	8006	38.31	26.76	57.39
Like	Aligned	6874	0.612	1.000	0.487
	Misaligned	8006	0.508	1.000	0.500
Dislike	Aligned	6874	0.310	0.000	0.462
	Misaligned	8006	0.395	0.000	0.489

Notes. This table reports summary statistics for dwell time and engagement outcomes by whether the perceived signal is aligned or misaligned with the participant's prior belief. An article is classified as "Aligned" if the perceived signal is on the same side of the neutral point (0.5) as the prior belief, and "Misaligned" otherwise.

neutral point (0.5) as their prior belief. Participants are more likely to "like" articles they perceive as aligned with their priors and are more likely to "dislike" articles they perceive as conflicting with their beliefs. We also see slightly higher mean and median dwell times for articles perceived to be misaligned. Given these correlations between alignment and engagement, we take these features into account when modeling engagement in Section 4.2.

4 A Model of Information Processing and Content Engagement

To decompose the respective roles of information-processing biases and recommendation algorithms in driving belief polarization, we develop a model of how individuals process information, update beliefs, and engage with recommended content. The model enables counterfactual simulations that quantify the respective roles of information-processing biases and recommendation algorithms in driving belief polarization.

We begin by describing the sequence of events that characterizes a content browsing session. We then identify two information-processing biases—*signal distortion* and *weight distortion*—using the standard Bayesian updating framework as a benchmark. Signal distortion refers to biased perceptions of the information. For example, confirmation bias can arise via a specific direction of signal distortion — by misinterpreting neutral or opposing information as more supportive of one's prior views (Rabin and Schrag, 1999). Weight distortion refers to ascribing different levels of credibility to new information based on its proximity to one's own beliefs, which subsequently

impacts the extent to which users update their beliefs when presented with new information (Mehta et al., 2008). The following section provides evidence for these two biases. We then model how these biases, together with content characteristics, shape engagement decisions—specifically, reading time and explicit feedback. Finally, we present estimates of the model.

We assume the following sequence of events. An individual i engages in a content browsing session in which a recommendation algorithm serves an article at each time period t. Let b_{it} be the prior belief at the beginning of period t, before exposure to the article.

1. **Reading stage**: Individual i spends time reading and processing the information. During this stage, they develop their interpretation of the article's information, i.e., signal.
2. **Belief updating stage**: Given their interpretation of the article's information and their prior belief, i updates their belief to their posterior belief.
3. **Feedback stage**: Individual i can leave feedback for the content, where they can "like", "dislike", or not react to the article.

This sequence of events reflects the experimental design, whereby the participant reads the article, reports their perceived signal and posterior belief in subsequent survey questions, and then chooses their feedback. We first focus on the belief updating stage, identifying biases in how individuals interpret and weight new information. We then develop models for the reading and feedback stages.

4.1 Identifying Information Processing Biases

We begin by presenting a benchmark model of how information impacts beliefs under Bayesian learning. During the belief updating stage, an individual is presented with one article. Because they see one article per time period, we denote both the article and time period by t.

1. An individual i has a prior belief b_{it} about the harmfulness of GMOs. We assume that this prior belief comes from a normal distribution, $b_{it} \sim N(\bar{b}_{it}, (\sigma^b_{it})^2)$.
2. In period t, individual i reads an article with a true signal s^T_{it} about the harmfulness of GMOs. We assume that the true signal comes from a normal distribution $s^T_{it} \sim N(\bar{S}_{it}, \sigma^2_{it})$. The individual forms an interpretation of this signal, denoted by s^P_{it}, which could be different from the true signal that they receive such that $s^P_{it} = s^T_{it} + \Delta_{it}$.
3. Given normal-normal conjugacy, the posterior belief of the individual i also follows a normal distribution where the mean and variance of the posterior belief, b'_{it} and $(\sigma^{b'}_{it})^2$, are:

$$b'_{it} = \left[\frac{(\sigma^S_{it})^2}{(\sigma^b_{it})^2 + (\sigma^S_{it})^2}\right] b_{it} + \left[\frac{(\sigma^b_{it})^2}{(\sigma^b_{it})^2 + (\sigma^S_{it})^2}\right] s^P_{it}, \tag{1}$$

$$(\sigma^{b'}_{it})^2 = \left(\frac{1}{(\sigma^b_{it})^2} + \frac{1}{(\sigma^S_{it})^2}\right)^{-1}. \tag{2}$$

Note that the variance of the perceived signal, $(\sigma^S_{it})^2$, determines the extent that i updates their prior towards the perceived signal. Larger signal variance, holding all else fixed, would result in smaller updates to beliefs based on new information. The individual repeats this process for ten periods such that the distribution of the posterior belief at period t is equivalent to the distribution of the prior belief at $t+1$.

The model presented above is a general version of the Bayesian learning model that is usually employed in empirical applications, especially when using observational data (Erdem and Keane, 1996; Iyengar et al., 2007; Narayanan and Manchanda, 2009; Sriram et al., 2015). In particular, there are two ways in which the above model extends to canonical Bayesian framework. First, in the absence of information on how recipients interpret the signal, researchers typically assume that all recipients interpret the signal in the same manner, i.e., $s^P_{it} = s^T_t \ \forall i$. We relax this assumption by allowing the signal's perception to vary across recipients. Second, it is common practice to assume that the perceived precision of the signal is invariant within a recipient over time, i.e., $(\sigma^S_{it})^2 = (\sigma^S_i)^2$ $\forall t$. In contrast, our model allows perceived precision to vary by the article content. This flexibility enables us to capture two distinct channels of bias in information processing — signal distortion and weight distortion — by allowing both the perceived signal and its perceived precision to depend on the context in which the information is received. Given that we observe how signal is perceived vis-a-vis the objective value and can recover how the perceived precision varies across individuals, we can explore if these deviations systematically vary as a function of some observable contextual variables.

To understand how information processing biases vary by signals and prior beliefs, we categorize signals into three mutually exclusive zones based on where the signal falls relative to an individual's prior belief and the neutral point (0.5 on our scale). Consider an individual with a pro-GMO prior (e.g., $b_{it} = 0.7$). A signal can fall into one of three zones:

1. More extreme, same direction: This signal aligns with the direction of the individual's prior, but also is more extreme (e.g., $s^T_{it} = 0.9$).

2. Less extreme, same direction: This signal confirms the individual's general stance but is more moderate (e.g., $s_{it}^T = 0.6$).

3. Opposite direction: The signal is on the opposite side of neutral from the prior (e.g., $s_{it}^T = 0.3$). This signal contradicts the individual's prior.

Comparing distortion across these zones allows us to examine how information-processing biases depend on the relationship between incoming signals and prior beliefs. For instance, confirmation bias predicts that individuals may systematically misinterpret signals that contradict their priors by shifting them toward previously held beliefs. Our framework enables us to test this prediction by directly measuring signal distortion as a function of whether the true signal confirms or contradicts the prior. We can also assess whether individuals distort either the signal itself or its perceived precision when confronted with signals that are belief-consistent but vary in extremeness. Below, we characterize patterns of signal distortion and weight distortion across these zones.

4.1.1 Signal Distortion

We investigate the presence of signal distortion by estimating the following general regression specification:

$$SignalDistortion_{it} = \psi_i + \psi_t + \psi_s + \psi_z \cdot (\text{Prior-Signal Diff}_{it} \times z_{it}) + \epsilon_{it}, \tag{3}$$

where the dependent variable is defined as $(s_{it}^P - s_{it}^T)$, which is the amount by which the true signal is misperceived by individual i in period t. The independent variables of interest are Prior-Signal Diff $:= (b_{it} - s_{it}^T)$ and how its effect differs by the zone, i.e., the direction of the signal relative to the prior (z_{it}). Specifically, z_{it} is an indicator for whether the perceived signal falls into one of the three mutually exclusive zones $z \in \{\text{More Extreme Same Dir}, \text{Less Extreme Same Dir}, \text{Opposite Dir}\}$. If the signal equals the prior, we classify that as the signal being less extreme in the same direction.[8] Recall that we (the researchers) observe the perceived and true signals, s_{it}^P and s_{it}^T, respectively, as well as the prior belief before observing the signal, b_{it}.[9] We control for individual-level tendencies in

[8] Such instances are rare. This occurs in 61 (0.04%) observations.

[9] s_{it}^P is collected by asking users about their perception of the article's stance on the harmfulness of GMOs. Since s_{it}^T is the intensity score specified in the GPT-4 prompt, it is measured on a bipolar scale where values near -1 indicate strong opposition to GMOs and values near 1 indicate strong support. We scale s_{it}^T from 0 to 1 using the transformation $\frac{x+1}{2}$ to ensure that all constructs are on a scale from 0 to 1. b_{it} is the posterior belief from the previous period which is also collected by asking users about their belief about the harmfulness of GMOs after they have read the article.

how information is interpreted by including individual fixed effects ψ_i. For instance, some individuals may consistently perceive articles as more favorable toward GMOs than the objective content would suggest. We also include time fixed effects ψ_t to control for common shocks to perception that vary across periods, such as fatigue over the course of the browsing session. Finally, because our focus is on how the signal is distorted by its proximity to the individual's prior, we include signal-bin fixed effects ψ_s, where each bin s corresponds to a discretized interval of the true signal $s_{it}^T \in [0, 1]$ of width 0.1. These fixed effects flexibly absorb misperceptions that are common across individuals exposed to articles whose anti- versus pro-GMO stance falls within a given bin (as seen in Figure 3). To measure heterogeneity in distortion across zones determined by the signal's location relative to the prior, we interact the prior–signal difference with zone-specific indicators.

The key parameters in Equation 3 are ψ_z. A positive value of ψ_z indicates distortion *toward* the prior, and a negative estimate indicates distortion *away from* the prior for that given zone z. To see this, let us first consider the main effect of the zones, especially as it applies to instances where the dis-confirmatory signals, i.e., signal is in the opposite direction of the prior belief. If the prior belief is anti-GMO (i.e., < 0.5), the signal would be pro-GMO (i.e., > 0.5). Thus, Prior-Signal Diff $:= (b_{it} - s_{it}^T < 0)$. Therefore, a positive ψ_z would imply a negative value of the product, thereby rendering the perceived signal to be smaller and pulled closer to the prior belief. In a similar vein, if the prior belief is > 0.5 and the signal is dis-confirmatory, i.e., < 0.5, Prior-Signal Diff $:= (b_{it} - s_{it}^T > 0)$. Therefore, a positive value ψ_z would render the product positive, pulling the perceived signal closer to the prior. Using this line of reasoning, we can verify that a positive value of ψ_z would pull the perceived signal closer to the prior belief for all three zones. Turning to the interaction effects of the zones with Prior-Signal Diff, when the true signal exceeds the prior ($s_{it}^T > b_{it}$), we have $b_{it} - s_{it}^T < 0$, so a positive ψ_1 reduces the perceived signal, pulling it back toward the prior. Conversely, when the true signal is below the prior ($s_{it}^T \leq b_{it}$), $b_{it} - s_{it}^T \geq 0$, a positive ψ_1 increases the perceived signal, again moving it toward the prior.

We report the estimates of Equation 3 in Table 3. Column (1) presents results without fixed effects. The intercept is 0.09, indicating that, on average, individuals perceive signals as more pro-GMO than they truly are. Although this positive intercept suggests a general upward bias in signal interpretation, our primary focus is on distortion that is systematically related to the position of the signal relative to the prior belief. These distortions are captured by the (a) the fixed effects corresponding to the three zones where the signal could fall relative to an individual's prior belief and (b) the coefficients on the prior–signal difference interacted with the indicators for the three

zones. Of these, (a) quantifies the extent to which individuals distort signals within each of the three zones. On the other hand, by allowing the effect of signal–prior proximity to vary across zones, (b) enables us to assess whether—and to what degree—the perceived signal is systematically pulled toward the individual's prior belief. For (a), we set less extreme signals in the same direction as the prior belief as the baseline. Across columns 1–4, we find the following results.

1. For signals that are more extreme in the same direction, the main effect is negative (significant for models (1)-(3), but not for (4)), suggesting that the signal is perceived to be closer to the true value. As discussed earlier, a negative estimate also suggests that the perceived signal is farther away from the prior than the true signal. The corresponding interaction with prior-signal difference is positive and significant. This implies that for signals that fall in this zone, individuals tend to distort them such that the perception is closer to the prior belief. Of the two, given the magnitude of the estimates, the latter effect is likely to be more dominant, especially for model (4).
2. For signals that are in the opposite direction, the main effect is negative (significant for models (1)-(3), but not for (4)). The corresponding interaction of the interaction with prior-signal difference not statistically significant for all models, except model (1). Therefore, we infer that there is no signal distortion for signals that are contrary to an individual's prior belief.
3. For signals that are less extreme on the same direction, the main effect is zero by default. The corresponding interaction with prior-signal difference is positive and significant for all models. Once again, this implies that individuals distort information such that it is perceived as being closer to their prior beliefs.

In summary, we find evidence of signal distortion towards an individual's prior belief when the signals are confirmatory, both for for highly extreme and more moderate signals. We the effects implied by the estimates in Figure 4. Thus, signal distortion is likely to attenuate belief updating in response to confirmatory signals, as such signals are perceived to lie closer to the individual's prior and therefore convey less incremental information.

4.1.2 Weight Distortion

Recall that we define weight distortion as deviations from the weight that an individual actually places on the new information (i.e., the signal) relative to what is implied under Bayesian updating.

Table 3: Evidence of Signal Distortion

	Signal Distortion			
	(1)	(2)	(3)	(4)
Constant	0.0907***			
	(0.0040)			
More Extreme, Same Dir	-0.0477***	-0.0453***	-0.0447***	-8.54×10^{-5}
	(0.0050)	(0.0054)	(0.0054)	(0.0057)
More Extreme, Same Dir × Prior-Signal Diff	0.2551***	0.2908***	0.2894***	0.1087***
	(0.0194)	(0.0213)	(0.0212)	(0.0257)
Opposite Dir	-0.0222***	-0.0204***	-0.0198***	0.0045
	(0.0050)	(0.0053)	(0.0053)	(0.0053)
Opposite Dir × Prior-Signal Diff	0.0335***	0.0095	0.0104	-0.0010
	(0.0078)	(0.0071)	(0.0071)	(0.0098)
Less Extreme, Same Dir × Prior-Signal Diff	0.5902***	0.5293***	0.5322***	0.1117***
	(0.0208)	(0.0248)	(0.0247)	(0.0285)
Observations	14,880	14,880	14,880	14,880
R^2	0.08847	0.23275	0.23496	0.35842
Participant fixed effects		✓	✓	✓
Period fixed effects			✓	✓
True Signal Bin fixed effects				✓

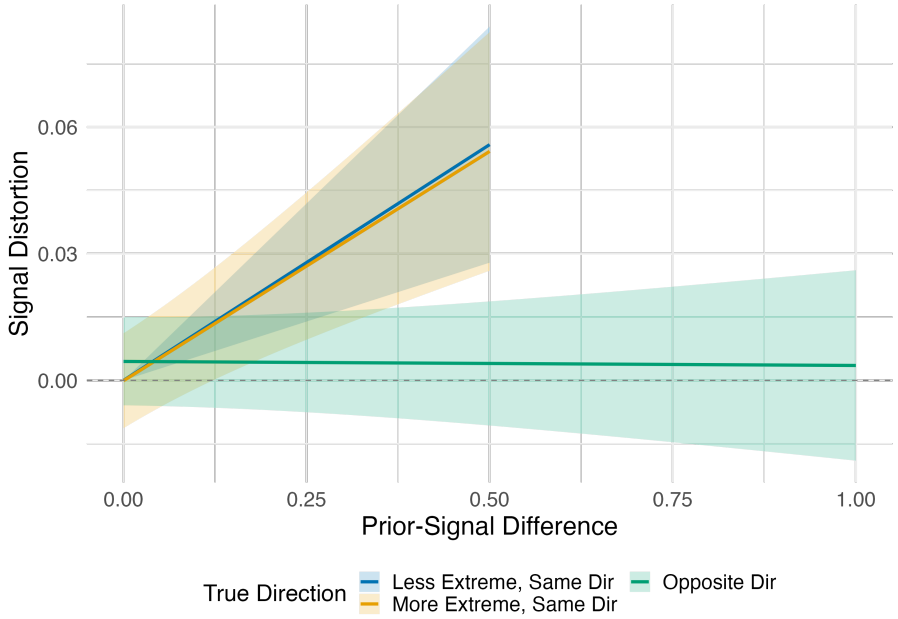
Notes. ***, **, * indicate significance at the 1%, 5%, and 10% level respectively. Standard errors are clustered at the individual level.

Given that we observe the signal as perceived by individual i at time t (s_{i1}^P), i's prior belief (b'_{it-1}), and i's posterior belief (b'_{it}), we can trivially back out the weight, w_{it}^{actual}, that the individual places on new information in each period from Equation 1. The intuition is that Equation 1 represents the posterior belief as a convex combination of the individual's prior belief and the perceived signal. Thus, we can compute w_{it}^{actual} for individual i in period t directly from the observed belief updates as $w_{it}^{actual} := \frac{b'_{it}-b_{it}}{s_{it}^P-b_{it}}$.[10]

In order to determine the benchmark weights that the individual should have placed under Bayesian updating, we rely on Equation 2. In this equation, if we know the variance of signal, σ_{it}^S, we can infer the posterior variance, σ_{it}^b, recursively. It would then be trivial to compute the weight that individual i should place on the signal they receive in period t, s_{it}^P, using Equation 1. To this end, we invoke the rational expectations assumption that is typically employed in Bayesian

[10]The Bayesian learning model also imposes the assumption that the posterior mean always lies between the convex hull of the prior mean and the perceived signal. In 22.81% of observations, there are minor violations of this assumption and in our main analysis, we censor the data to fulfill this assumption. We also plan to conduct a robustness check where we accommodate violations of this assumption through a "clicking error" while reporting posterior beliefs on the slider scale as shown in Figure A.2.

Figure 4: Signal Distortion Implied by the Estimates



Notes. The figure illustrates that for signals that are confirmatory (both extreme and moderate), the distortion increases with the distance of the prior belief relative to the signal. Thus, individuals receiving confirmatory signals are likely to perceive the signals as being closer to their prior beliefs.

updating models (e.g., Erdem and Keane, 1996) and fix individual i's benchmark signal variance as the empirical variance of the true signals in our content repository. Therefore, σ_{it}^S remains constant for all individuals and does not vary based on the alignment of the signal with their prior belief.

Our primary interest is in systematic forms of distortion, specifically in how the weight placed on the signal varies with its perceived distance from the prior belief. We examine this by estimating the following linear regression:

$$w_{it}^{actual} - w_{it}^{benchmark} = \kappa_i + \kappa_t + \kappa_s + \kappa_z \cdot (\text{DistToPrior}_{it} \times z_{it}) + \epsilon_{it}. \tag{4}$$

The dependent variable, $w_{it}^{actual} - w_{it}^{benchmark}$ captures the extent to which the actual weight placed on the signal at time t, w_{it}^{actual}, differs differs from the weight implied by the distortion-free Bayesian benchmark. Positive values indicate over-weighting and negative values indicate under-weighting relative to the benchmark. As with the signal distortion specification, z_{it} is an indicator for whether the perceived signal falls into zone $z \in \{\text{More Extreme Same Dir}, \text{Less Extreme Same Dir}, \text{Opposite Dir}\}$,

and $DistToPrior_{it} = |s_{it}^P - b_{it}|$ is the distance between the perceived signal and the prior. We focus on the perceived signal, instead of the true signal, to rule out the possibility that any bias in updating is due to bias in signal interpretation. The coefficients κ_z capture how the weight deviation varies with distance in each zone. A positive estimate of κ_z means individuals update more, relative to the distortion-free benchmark, when perceived signals that are farther from the prior. We include participant fixed effects (κ_i), period fixed effects (κ_t), true signal bin fixed effects (κ_s), and zone fixed effects (κ_z) progressively across specifications.

Table 4: Evidence of Weight Distortion

	Actual - Benchmark Weight			
	(1)	(2)	(3)	(4)
Constant	0.4921***			
	(0.0151)			
More Extreme, Same Dir	-0.3553***	-0.3630***	-0.3604***	-0.3540***
	(0.0215)	(0.0222)	(0.0215)	(0.0214)
More Extreme, Same Dir × Distance to Prior	-0.0833	0.1138*	0.1536***	0.1967***
	(0.0596)	(0.0601)	(0.0554)	(0.0556)
Opposite Dir	-0.2034***	-0.1176***	-0.1598***	-0.1828***
	(0.0222)	(0.0202)	(0.0194)	(0.0197)
Opposite Dir × Distance to Prior	-0.0284	-0.1248***	-0.0204	0.0277
	(0.0300)	(0.0231)	(0.0227)	(0.0239)
Less Extreme, Same Dir × Distance to Prior	-0.7853***	-0.8319***	-0.6782***	-0.7483***
	(0.0971)	(0.0915)	(0.0928)	(0.0928)
Observations	14,880	14,880	14,880	14,880
R^2	0.05883	0.36185	0.45702	0.46323
Participant fixed effects		✓	✓	✓
Period fixed effects			✓	✓
True Signal Bin fixed effects				✓

Notes. ***, **, * indicate significance at the 1%, 5%, and 10% level respectively. All columns present OLS estimates of Equation 4. Standard errors are clustered at the individual level.

We report the estimates of Equation 4 in Table 4. Column (1) presents estimates without fixed effects. The positive constant (0.4921) indicates that, on average, individuals place more weight on the signal than the Bayesian benchmark would predict. Nevertheless, our primary objective is to assess whether individuals systematically over- or under-weight new information—relative to the Bayesian benchmark—depending on the signal's position with respect to their prior belief. The results in columns 1–4 reveal the the following patterns of weight distortion across the three zones.

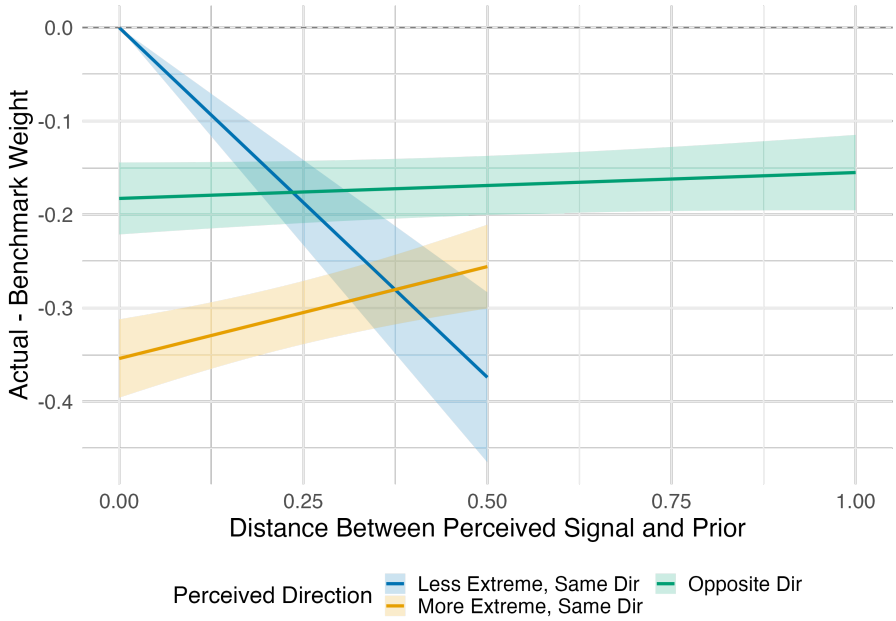
1. For signals that are more extreme in the same direction, the main effect is negative and

statistically significant, suggesting that individuals place a lower weight on new information that falls in this zone. The corresponding interaction with the distance to the prior is positive and significant. This implies that, within this zone, the degree of underweighting diminishes as the signal moves farther from the prior belief—that is, individuals discount new information less when it is more distant from their prior.

2. For signals that are in the opposite direction, the main effect is negative and statistically significant, suggesting that individuals tend to underweight non-confirmatory new information. The corresponding interaction of the interaction with distance to the prior is not statistically significant for all models, except model (2). Therefore, we infer that there is no weight distortion as a function of proximity for signals that are contrary to an individual's prior belief.

3. For signals that are less extreme in the same direction, the main effect is zero by default. The corresponding interaction with the distance to the prior is negative and significant for all models. This implies that, within this zone, individuals place relatively less weight on new information the farther it lies from their prior belief.

Figure 5: Weight Distortion Implied by the Estimates



Notes. The figure illustrates that actual weight on new information is lower than the benchmark Bayesian weights for all zones.

We present a visual representation of these results in Figure 5. [11] Overall, the results suggest a negative weight distortion across the board. Therefore, individuals tend to underweight new information relative to what we would expect under the benchmark Bayesian updating model. Thus, individuals may be less open to new information when updating their beliefs.

4.2 A Model of Engagement

Having established evidence for signal and weight distortion in the previous section, we now develop a structural model that captures how individuals make engagement decisions in response to recommended content.

Reading stage. Empirically, dwell time is measured as the duration for which the article remains open on the individual's screen. Thus, this measure captures not only reading time, but also the time spent thinking about and processing the information. Therefore, in this stage, the individual forms their perception of the article's signal s_{it}^P. We assume that individual i chooses a dwell time τ_{it} to maximize the following per-period payoff:

$$U(\tau_{it}) = a(x_{it}^D)\, \eta_{it}\, g(\tau_{it}) - C_i(\tau_{it}, N_{it}). \tag{5}$$

The utility function decomposes into (i) a benefit from allocating attention to the article and (ii) the opportunity cost of time. This structure formalizes the idea that reading time is chosen to equate the marginal value of consuming additional content with the marginal cost of time, allowing engagement to vary systematically with both article characteristics and the individual's beliefs.

The benefit term $a(x_{it}^D)\, \eta_{it}\, g(\tau_{it})$ captures the total "effective attention" allocated to the article. We model effective attention as the product of three components. First, $a(x_{it}^D)$ governs the baseline intensity with which the article can attract and sustain attention and is a function of the attributes of the article and the individual at the time of exposure, i.e., x_{it}^D. Second, $\eta_{it} > 0$ is an attention productivity shock that captures transitory fluctuations in the individual's ability to convert time into attention (e.g., distraction, multitasking, or device frictions). Third, $g(\tau_{it})$ maps dwell time into accumulated attention and is assumed to be increasing and concave, reflecting diminishing marginal

[11] Note that the range of the X-axis for the different zones is determined by the values the distance variables can take. For confirmatory signals (both extreme and moderate), the maximum distance of the signal from the prior is 0.5. However, for non-confirmatory signals, the distance can be as high as 1.

returns to spending additional time reading. We parameterize attention intensity as:

$$a(x_{it}^D) = \exp\big(\delta_0 + \Delta' x_{it}^D\big)\,, \tag{6}$$

which ensures $a(x_{it}^D) > 0$ and allows covariates to shift attention intensity proportionally. The covariate vector x_{it}^D includes the distance to the prior and its interaction with each zone, similar to Equation 4. In addition, we include the length of the article length (the number of characters in the article) because it is likely to have a direct impact the dwell time.[12]

The multiplicative term $\eta_{it} > 0$ is an attention productivity shock that captures transitory factors affecting engagement conditional on content, such as distractions, multitasking, device frictions, or momentary fluctuations in cognitive focus. We assume η_{it} is realized after the article is displayed but before the dwell-time decision, so individuals optimally condition reading time on its realization. We further assume that η_{it} is i.i.d.

The function $g(\tau_{it})$ maps reading time into attention. To capture diminishing returns to additional time spent reading, we specify:

$$g(\tau_{it}) = \log(1 + \tau_{it}). \tag{7}$$

The second term, $C_i(\tau, N_{it})$, represents the opportunity cost of time and is assumed to increase with session fatigue. In particular, we assume the per-second cost of time is strictly positive and depends on the number of articles already read in the session:

$$C_i(\tau_{it}, N_{it}) = \exp(c_i + f(N_{it}))\ \tau_{it}, \qquad c_i \in \mathbb{R},\ \gamma \geq 0. \tag{8}$$

Here, c_i captures individual-specific baseline opportunity costs, while N_{it} denotes the number of articles that individual i has already read in the session. The parameter γ governs how fatigue accumulates over the session by scaling the marginal opportunity cost of time. Given this utility specification, individual i chooses dwell time τ_{it} to maximize $U(\tau_{it})$:

$$\tau_{it} \in \arg\max_{\tau_{it}>0}\ \left\{a(x_{it}^D)\,\eta_{it}\ \log(1+\tau_{it}) - \exp(c_i + f(N_{it}))\,\tau_{it}\right\}. \tag{9}$$

In Appendix C.1, we show how solving this decision problem results in an estimation equation as

[12] We standardize article length across all articles by computing the z-score.

follows:

$$\log(1+\tau_{it}) = \alpha_i + \Delta' x_{it}^D - f(N_{it}) + \epsilon_{it}. \tag{10}$$

where

$$\alpha_i = \delta_0 - c_i.$$

Belief updating stage. In this stage, individual i uses the perceived signal to update their posterior belief about the harmfulness of GMOs. We assume that they update their beliefs using the Bayesian learning framework described in Section 4.1, which allows for signal distortion through their perceived signal and weight distortion through flexibly estimating the signal variance.

Feedback stage. After updating beliefs, individual i chooses an explicit reaction to the article—"dislike", "no reaction", or "like". We assume that this choice is driven by the utility the individual derives from the article's content as interpreted after reading it, and does not depend on dwell time, which is a sunk cost at this stage. We model the reaction decision through a latent content-utility index:

$$u_{it} \;\equiv\; u(x_{it}^F) = \beta' x_{it}^F + \xi_{it}, \tag{11}$$

where x_{it}^F are attributes that affect an article's utility and ξ_{it} is an idiosyncratic taste shock capturing unobserved factors affecting expressed reactions. This specification formalizes the idea that reactions are discrete expressions of favorability: individuals map the latent utility from the interpreted content into one of three ordered responses. We assume that individuals "like" articles with greater utility and "dislike" articles with lower utility.

Let $y_{it} \in \{-1, 0, 1\}$ denote the observed reaction, where $y_{it} = -1$ corresponds to "dislike", $y_{it} = 0$ to *no reaction*, and $y_{it} = 1$ to "like". We model choices via two thresholds $\delta_d < \delta_l$ such that:

$$y_{it} = \begin{cases} -1 & \text{if } u_{it} < \delta_d, \\ 0 & \text{if } \delta_d \le u_{it} \le \delta_l, \\ 1 & \text{if } u_{it} > \delta_l. \end{cases} \tag{12}$$

Further, we assume that ξ_{it} follows a logistic distribution with CDF as follows:

$$\Lambda(v) \;\equiv\; \Pr(\xi_{it} \le v) = \frac{1}{1+e^{-v}}. \tag{13}$$

In Appendix C.2, we derive the following estimation equation resulting from our model:

$$\log\left(\frac{\Pr(y_{it} \leq k \mid x_{it}^F)}{1 - \Pr(y_{it} \leq k \mid x_{it}^F)}\right) = \delta_k - \beta x_{it}^F, \qquad k \in \{-1, 0\}. \tag{14}$$

Equation (14) is the ordered logit model: the slope on x_{it}^F is common across cutpoints (the proportional-odds restriction), while the intercept varies across cutoffs via δ_k. Table 2 shows that "likes" and "dislikes" correlate with how aligned the content is with the individual's prior belief. Therefore, we use the same covariates in x_{it}^F as in x_{it}^D.

5 Model Estimates and Simulation Method

We present the estimates of the model of dwell time in Table 5. As expected, the results imply that individuals spend more time reading longer articles. Moreover, individuals spend less time reading as the session progresses, consistent with session fatigue.[13] The results are robust to the inclusion of period and signal bin fixed effects (Columns 2 and 3).

We now turn to the main results linking dwell time to the position of new information relative to the individual's prior belief. As in the case of signal and weight distortion, we present the results for the three zones defined based on the position of the signal relative to the prior belief.

1. For signals that are more extreme in the same direction, the main effect is not statistically significant. The corresponding interaction with the distance to the prior is negative for all models and significant for model (2). This implies that within this zone, individuals spend less time reading articles that are farther from their prior belief.

2. For signals that are in the opposite direction, the main effect is positive and statistically significant, suggesting that individuals tend to spend more time reading non-confirmatory new information. The corresponding interaction of the interaction with the distance to the prior is negative and statistically significant for all models. These results present an interesting tussle between the positive main effect of the zone and the negative effect of distance to the prior within this zone. The net effect is positive across most values of distance. The primary exception arises for individuals with extreme prior beliefs who encounter highly disconfirmatory signals; in these cases, the distance measure can become sufficiently large to

[13] Within the ten time periods in our data, the negative linear effect is greater than the positive quadratic effect.

Table 5: Dwell Time Model Estimates

	Log(Dwell Time)		
	(1)	(2)	(3)
More Extreme, Same Dir	0.0429	0.0442	0.0474
	(0.0301)	(0.0301)	(0.0301)
More Extreme, Same Dir × Distance to Prior	-0.1363	-0.1549*	-0.1218
	(0.0929)	(0.0928)	(0.0928)
Opposite Dir	0.0806***	0.0818***	0.0704**
	(0.0296)	(0.0296)	(0.0301)
Opposite Dir × Distance to Prior	-0.0861**	-0.0967***	-0.0704*
	(0.0368)	(0.0369)	(0.0385)
Less Extreme, Same Dir × Distance to Prior	0.5007***	0.4675***	0.4319***
	(0.1551)	(0.1546)	(0.1555)
Article Length	0.1291***	0.1293***	0.1387***
	(0.0057)	(0.0057)	(0.0082)
Period	-0.1596***		
	(0.0088)		
Period square	0.0085***		
	(0.0008)		
Observations	14,880	14,880	14,880
R^2	0.62114	0.62342	0.62413
Participant fixed effects	✓	✓	✓
Period fixed effects		✓	✓
True Signal Bin fixed effects			✓

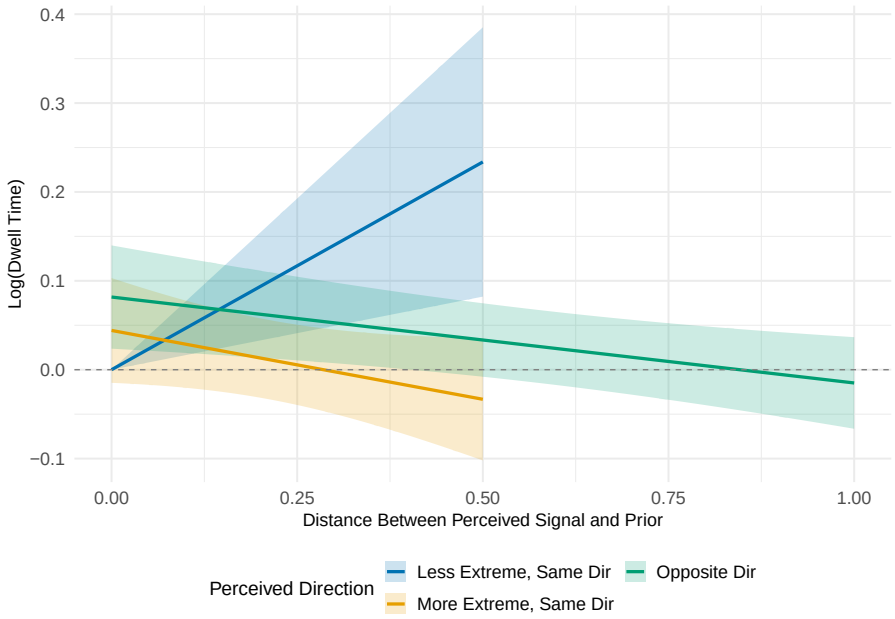
Notes. ***, **, * indicate significance at the 1%, 5%, and 10% level respectively. Standard errors are clustered at the individual level.

generate a net negative effect.

3. For signals that are less extreme in the same direction, the main effect is zero by default. The corresponding interaction with the distance to the prior is positive and significant for all models. This implies that, within this zone, dwell time is higher for articles that are far from the individual's prior belief.

We present a visual representation of these results in Figure 6. These results present an interesting tension between dis-confirmatory articles (i.e., (2) above) and moderate articles that are in the same direction as the prior belief. For moderate individuals, whose prior belief is close to 0.5 on either side, articles in the opposite direction may result in a higher dwell time compared to less extreme articles in the same direction. This is because the latter articles cannot be sufficiently be far away from the prior to benefit from the positive effect of distance on dwell time. On the other hand, for individuals with extreme prior beliefs, less moderate articles in the same direction

Figure 6: Effect on Dwell Time Implied by the Estimates



Notes. These results present an interesting tension between dis-confirmatory articles (i.e., (2) above) and moderate articles that are in the same direction as the prior belief. For moderate individuals, whose prior belief is close to 0.5 on either side, articles in the opposite direction may result in a higher dwell time compared to less extreme articles in the same direction. On the other hand, for individuals with extreme prior beliefs, less moderate articles in the same direction are likely to have higher dwell time.

are likely to have higher dwell time. Therefore, an algorithm that maximizes dwell time might feed different types of content based on where the individual's prior belief is located (even though it does not know this prior belief).

Next, we present the estimates for feedback in Table 6. The results indicate that individuals are less likely to express positive engagement with longer articles. Moreover, individuals spend less time reading as the session progresses.[14] The results are robust to the inclusion of period fixed effects (Column 3). We now turn to the main results linking feedback provision to the position of new information relative to the individual's prior belief. As in the case of signal and weight distortion as well as dwell time, we present the results for the three zones defined based on the position of the signal relative to the prior belief.

1. For signals that are more extreme in the same direction, the main effect is not statistically

[14]Within the ten time periods in our data, the negative linear effect is greater than the positive quadratic effect.

Table 6: Feedback Model Estimates

	Cumulative log-odds		
	(1)	(2)	(3)
More Extreme, Same Dir	−0.042	−0.043	−0.055
	(0.080)	(0.080)	(0.080)
More Extreme, Same Dir × Distance to Prior	−1.307***	−1.304***	−1.203***
	(0.249)	(0.249)	(0.251)
Opposite Dir	0.071	0.073	0.039
	(0.077)	(0.077)	(0.078)
Opposite Dir × Distance to Prior	−1.325***	−1.333***	−1.253***
	(0.095)	(0.095)	(0.098)
Less Extreme, Same Dir × Distance to Prior	−0.149	−0.170	−0.269
	(0.414)	(0.415)	(0.417)
Article Length	−0.056***	−0.057***	0.049**
	(0.016)	(0.017)	(0.024)
NoReact Threshold	−1.427***	−1.387***	−0.963***
	(0.084)	(0.078)	(0.095)
Like Threshold	−1.047***	−1.007***	−0.581***
	(0.083)	(0.076)	(0.095)
Period	−0.135***		
	(0.026)		
Period square	0.009***		
	(0.002)		
Observations	14 880	14 880	14 880
AIC	26 559.96	26 563.31	26 508.77
Period fixed effects		✓	✓
True Signal Bin fixed effects			✓

Notes. ***, **, * indicate significance at the 1%, 5%, and 10% level respectively. This table presents the ordered logit model estimates in Equation 14. The dependent variable is $\log\left(\frac{\Pr(y_{it}\leq k|x_{it}^F)}{1-\Pr(y_{it}\leq k|x_{it}^F)}\right)$. Negative values indicate a shift toward negative reactions (from "like" to no reaction to "dislike"). Standard errors are clustered at the individual level.

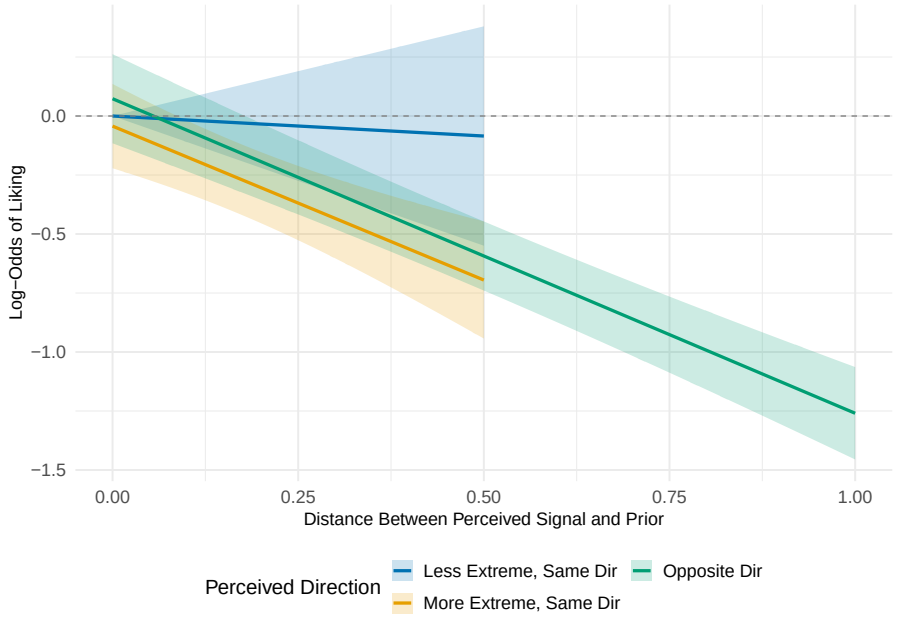
significant. The corresponding interaction with the distance to the prior is negative and statistically significant for all models. This implies that within this zone, individuals are less likely to "like" articles that are farther from their prior belief.

2. For signals that are in the opposite direction, the main effect is not statistically significant. The corresponding interaction of the interaction with the distance to the prior is negative and statistically significant for all models. Thus, individuals are less likely to "like" articles that are farther away from their prior belief within this zone.

3. For signals that are less extreme in the same direction, the main effect is zero by default. The corresponding interaction with the distance to the prior is not statistically significant for all

models.

Figure 7: Effect on Feedback Implied by the Estimates



Notes. These results suggest that articles that are distant from an individual's prior belief are less likely to receive positive engagement if they are dis-confirmatory or more extreme in the same direction.

We present a visual representation of these results in Figure 7. The key insight is that articles that are distant from an individual's prior belief are less likely to receive positive engagement if they are dis-confirmatory or more extreme in the same direction. Therefore, an algorithm that is trying to maximize "likes" will be inclined to serve articles that are closely aligned with an individual's prior belief.

Table 7 summarizes the directions of the effects of signal distortion, weight distortion, dwell time, and feedback for each zone (relative to the prior belief). We find evidence of signal distortion to the prior for both extreme and moderate confirmatory content. This kind of distortion implies that it may be harder to move individuals away from their prior when they see confirming signals. On the other hand, we do not find evidence that individuals systematically misinterpret disconfirmatory signals (i.e., contrary to their prior beliefs) as confirmatory. Thus, disconfirmatory information is likely to be interpreted objectively. We also find that individuals tend to put less weight on new information relative to an unbiased benchmark. This is particularly true for dis-confirmatory and for more

extreme confirmatory information. Taken together, these findings indicate that individuals exhibit a general reluctance to update their beliefs. While resistance to dis-confirmatory information can exacerbate polarization, a similar lack of responsiveness to more extreme confirmatory information is likely to yield more nuanced effects.

In addition, our findings on engagement highlight an important distinction between dwell time and feedback, two engagement metrics that recommendation algorithms may seek to maximize. Content that is "liked" does not necessarily generate longer reading times, and content that holds attention does not necessarily receive more positive feedback. In particular, participants devote more time to content that is disconfirmatory and more moderate. However, they do not particularly "like" such content. Conversely, participants dislike content that is more extreme and disconfirmatory.

This divergence has potential implications for recommendation design. A recommendation algorithm that maximizes dwell time will tend to select different content than one that maximizes likes. In practice, a dwell-time–maximizing policy will recommend content that is far away from an individual's beliefs. A like-maximizing policy, by contrast, will avoid content that differs from the individual's prior. Exposure to these different types of content can have distinct implications for posterior beliefs and polarization, which we explore in the counterfactual simulations.

Table 7: Direction of Effects Relative to Less Extreme, Confirmatory Information

Zone	**Signal Distortion**	**Weight Distortion**	**Dwell Time**	**Pr(Like)**
More extreme, same dir.	$\rightarrow$ prior	$\downarrow$ weight	—	$\downarrow$
Opposite direction	—	$\downarrow$ weight	$\uparrow$	$\downarrow$

Notes. This table summarizes Tables 3 – 6. $\rightarrow$ prior = perceived signal distorted toward prior; $\downarrow/\uparrow$ weight = under-/over-update vs. Bayesian benchmark; $\downarrow/\uparrow$ dwell = less/more time relative to if the perceived signal is equal to the prior.; $\downarrow/\uparrow$ Pr(Like) = less/more likely to "like" relative to if the perceived signal is equal to the prior. "—" = no significant effect.

5.1 Simulation Method

Next, we assess the extent to which personalized engagement-optimizing recommendation systems influence the polarization of beliefs, and how this depends on whether the platform optimizes likes or dwell time. We also decompose the relative contributions of information processing biases and algorithmic content curation to any observed changes in polarization.

The key structural assumption in our counterfactual simulations is that each individual's behavioral parameters are invariant to the platform's choice of engagement metric. For example, we

assume that ψ_z—the zone-specific tendency to pull distant signals toward the prior—does not change depending on whether the platform optimizes likes or dwell time.

5.1.1 Training Deep Reinforcement Learning-Based Recommenders

To conduct this assessment, we train deep reinforcement learning (DRL)-based recommendation agents using the estimated structural model as a simulation environment. As discussed in Section 2, modern recommendation systems increasingly rely on DRL methods to learn recommendation policies from sequential user interactions (Chen et al., 2023). DRL is necessary in our setting because the agent must select 10 articles sequentially from a repository of 200, conditioning each choice on its current state—the history of articles recommended and engagement outcomes received up to that period. Traditional reinforcement learning methods such as Q-learning require maintaining a Q-value for every (state, action) pair. At each period, the agent chooses from up to 200 articles, and the number of possible states—distinct histories of articles recommended and engagement outcomes received—grows combinatorially with each period, making a tabular approach computationally infeasible. A Deep Q-Network (DQN) (Mnih et al., 2015; Van Hasselt et al., 2016; Song et al., 2021; Khurana et al., 2024) addresses this by using a neural network to approximate Q-values over the state space, enabling the agent to generalize across states without enumerating them. Since recommendation systems are often personalized, we train one DQN per participant to maximize engagement over a 10-period content browsing session (an "episode").

Reward Function

The reward function operationalizes the platform's engagement objective. For like-maximization, the reward at each period is drawn from a multinomial distribution over the three feedback outcomes (dislike, no reaction, like), with probabilities predicted by the estimated ordered logit model. A like yields a reward of 1 and other outcomes yield 0. For time-maximization, the reward is drawn from a normal distribution centered at the log dwell time predicted by the estimated dwell time model. The agent seeks to maximize cumulative reward over the episode. Crucially, the agent does not observe beliefs—its state consists only of the history of articles recommended and engagement outcomes received. Beliefs evolve in the background: they shape engagement, which shapes recommendations, which in turn shapes beliefs. Specifying such a closed-loop system enables us to infer the relative impact of personalized engagement-optimizing recommendation algorithms vs. biases in information

processing on belief polarization.

Convergence Criterion

We establish a baseline engagement level by simulating 100 episodes with random recommendations. The average cumulative rewards across the 100 episodes under the random recommendation policy serve as the baseline engagement level of that participant in our sample. It represents the level of engagement that the participant would provide absent any personalization. The baseline engagement level of each participant is used determine the when that participant's DRL agent will stop training. In particular, training proceeds until the agent's rolling mean reward over the last 30 episodes exceeds this baseline by at least 5%.[15]

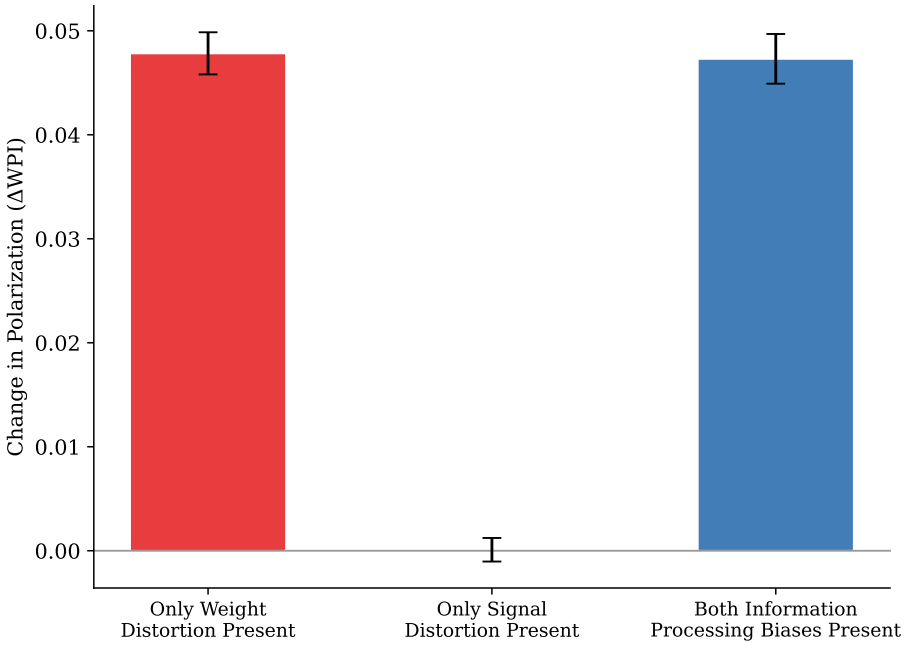
6 Decomposing the Drivers of Polarization

Finally, we advance to step 4 in our proposed methodology. Each trained agent is deployed over $K_{eval} = 100$ evaluation episodes. In each episode, the agent sequentially recommends 10 articles, selecting at each period the article it predicts will generate the highest engagement among those not yet recommended. After each recommendation, the participant's beliefs are updated through the structural model and an engagement outcome is realized, feeding back into the agent's next recommendation. To ensure comparability, the first article in each episode is fixed to match the corresponding random baseline episode.

Because the starting article varies across episodes and engagement outcomes are stochastic—both of which alter the agent's state and subsequent recommendations—belief trajectories differ across episodes even for the same participant. We therefore characterize each participant i by their average ending belief, $\overline{EndingBelief}_i = \frac{1}{K_{eval}} \sum_{e=1}^{K_{eval}} EndingBelief_{ie}$, and measure polarization in the distribution of $(\overline{EndingBelief}_i)_{i=1}^{n}$ using the Wolfson polarization index (WPI) (Wolfson, 1994, 1997; Wang and Tsui, 2000).

[15] A/B tests typically show mid-single-digit engagement lifts from personalization. For example, Pinterest reported ∼6% and 5% repin gains from successive feed personalization improvements (Xia et al., 2022). We treat 5% as a meaningful improvement bar.

Figure 8: Effect of Signal and Weight Distortion on Polarization



Notes. This figure plots the change in the Wolfson Polarization Index relative to the scenario in which both information processing biases are absent, under random recommendation. Error bars represent 95% bootstrapped confidence intervals.

6.1 Information Processing Biases

We first explore the role of signal and weight distortion in contributing to polarization by comparing polarization levels under the random recommendation algorithm when both biases are absent, only weight distortion is present, only signal distortion is present, and both biases are present. To simulate user behavior the absence of signal distortion, we set the perceived signal, s_{it}^P to be the true signal, s_{it}^T. To turn off weight distortion, we fix each individual's signal variance to be equal to the empirical distribution of the true signals, which equals 0.083 because the true signals are generated from a uniform distribution between 0 and 1.

Our simulations indicate that, in the absence of both biases, a recommendation algorithm that serves content uniformly at random reduces the WPI by 0.2664 relative to its initial level.[16] The reduction in polarization is a result of individuals being exposed to opposing points of view. Figure 8

[16]This value comes from comparing the WPI of the posterior distribution of beliefs at the end of the 10^{th} period and prior distribution of beliefs at the beginning of the 1^{st} period under this algorithm and without biases.

reports the change in polarization for alternative scenarios relative to this benchmark. We find that polarization declines under these counterfactual scenarios; however, introducing information-processing biases attenuates this reduction relative to the unbiased baseline. Accordingly, when both signal and weight distortions are present, the level of polarization after ten periods exceeds that observed in the absence of these biases. This pattern is evident in the third column, which shows a positive and statistically significant effect on the WPI when both signal and weight distortions are present, relative to the baseline. Decomposing this effect reveals that the increase in polarization is driven entirely by weight distortion: the estimates in the first and second columns indicate that signal distortion alone has little impact, whereas weight distortion drives the increase in polarization.

The null effect of signal distortion on polarization arises because signal distortion pulls perceived signals toward an individual's prior belief in a symmetric way, as shown in Table 3. True signals that are less extreme than the prior are perceived as more aligned with the prior, which means signals are perceived to be more extreme than they truly are. This exerts a polarizing force. At the same time, true signals that are more extreme than the prior are perceived as less extreme, which exerts a moderating force. As a result, these opposing adjustments offset one another, causing the net polarizing effect of signal distortion to cancel out.

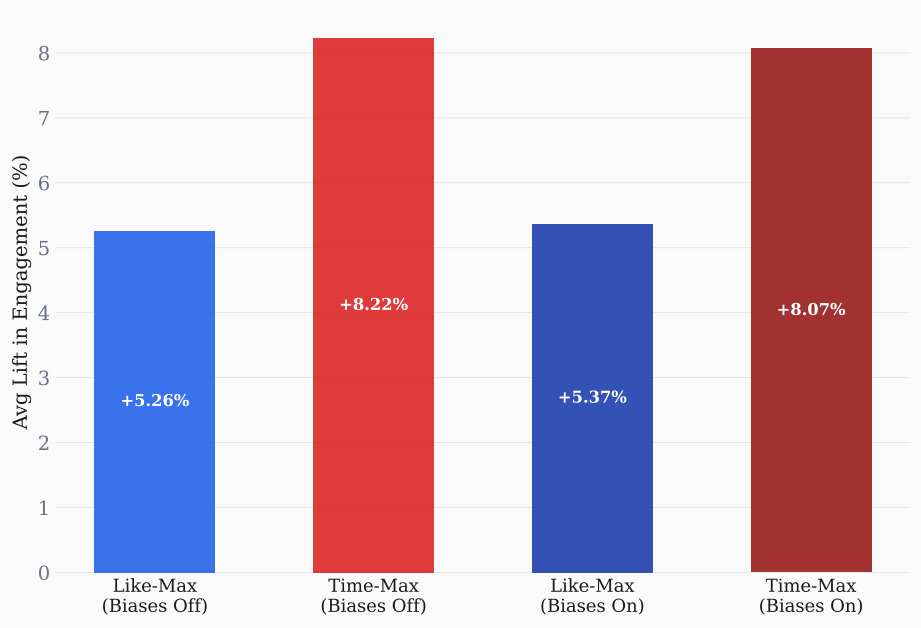
In contrast, weight distortion increases polarization because it leads people to put less weight on information that differs from their prior (see Table 4). In particular, individuals place less weight on signals that oppose their prior or that are more moderate but still distant from it (e.g., a signal of 0.5 when the prior is 1) or that are more extreme but confirmatory, relative to moderate information close to their prior. By systematically discounting disconfirming or distant information, weight distortion limits belief updating toward the center and instead reinforces prior beliefs.

6.2 Engagement-Maximizing Algorithms

In this section, we quantify the incremental polarizing effect of engagement-maximizing algorithms relative to a baseline in which content is presented uniformly at random. We consider two recommendation algorithms that differ in terms of their objective function: increasing "likes" vs. dwell time. We evaluate polarization under regimes with and without the information-processing biases. This approach allows us to verify that our results are robust to the inclusion of these biases and to rule out the possibility that the observed effects are driven by interactions between the distortions and the recommendation algorithms. As a sanity check, we assess whether the algorithms were

successful in increase the metrics they were meant to optimize. We present the results in Figure 9. Consistent with their objectives of increasing the engagement metrics by 5%, the like-maximizing and dwell-time–maximizing algorithms increase their respective engagement metrics relative to the baseline.

Figure 9: Lifts in Engagement



Notes. This figure plots the average lift in engagement across different scenarios. Specifically, it illustrates that DRL agents lift engagement in different scenarios relative to the baseline scenario where signals are recommended randomly and information processing biases are absent.

In Table 8, we report the change in the WPI relative to the baseline scenario where information processing biases are absent and content is recommended randomly. We report the changes in percentage terms in Table 8. We find that personalized recommendation systems optimized for engagement, such as like-maximization and dwell time-maximization, lead to greater belief polarization, both in the presence and absence of information processing biases. In order to understand why this happens, we plot the distribution of the distance between the true signal and the prior—absent any information biases—under three exposure regimes: random exposure, a like-maximizing algorithm, and a dwell-time-maximizing algorithm in Figure 10a.[17] The distribution of the distance

[17]We focus on the case without information biases because the interpretation of the perceived signal is more transparent—by construction, it coincides with the true signal. The qualitative patterns are similar when biases are present.

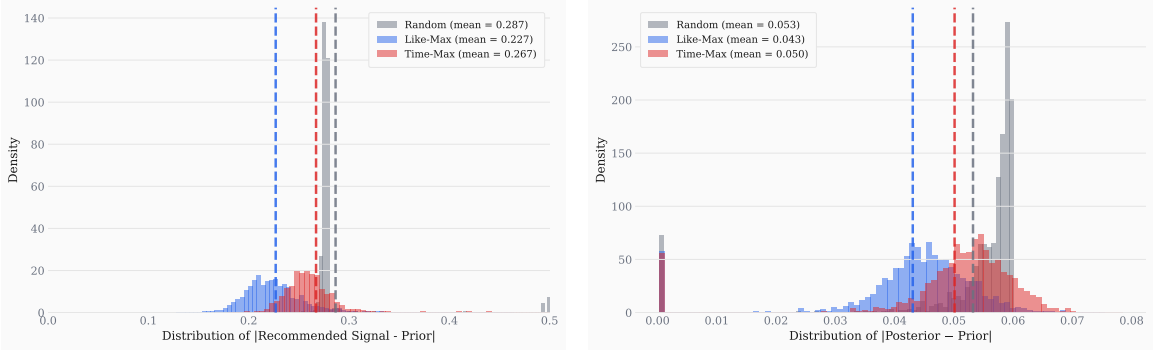
Table 8: Change in Polarization by Algorithms and Information Processing Biases

Scenario	**Random**	**Like**	**Time**
Information processing biases present	+17.76% (2.08%)	+28.84% (3.62%)	+22.18% (2.77%)
Information processing biases absent	–	+17.04% (2.19%)	+9.73% (1.30%)

Note: All scenarios exhibit depolarization relative to initial beliefs at period 0. This table reports the percentage by which each scenario reduces this depolarization relative to the baseline (no information processing biases, random recommendations) scenario. Specifically, each entry reports $(\Delta\text{WPI}_{\text{scenario}} - \Delta\text{WPI}_{\text{baseline}})/|\Delta\text{WPI}_{\text{baseline}}|\times 100$, where $\Delta\text{WPI} = \text{WPI}(t{=}10) - \text{WPI}(t{=}0)$. Positive values indicate that the scenario retains more polarization than the baseline. Bootstrapped standard errors in parentheses.

between the signal and the prior is more concentrated near zero under the like-maximizing algorithm than under the dwell-time-maximizing algorithm, while the random algorithm generates, on average, the greatest distance between the signal and the prior. Consequently, because individuals exposed to the random algorithm encounter more dissimilar information, they undergo larger belief updates. This pattern is reflected in Figure 10b, which shows greater posterior–prior distances under random recommendations relative to the engagement-maximizing regimes.

Figure 10: Content Exposure and Belief Updates Across Evaluation Episodes



(a) |Signal − Prior| (b) |Posterior − Prior|

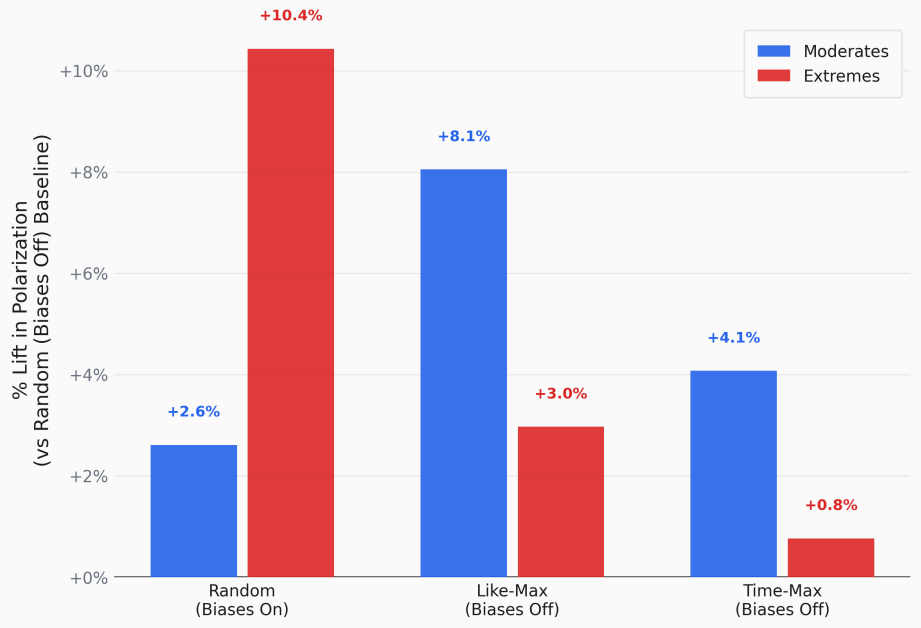
Notes. This figure plots the histogram of the mean absolute distance between recommended true signals and prior beliefs per user (panel a) and between the posterior and prior beliefs (panel b), across all periods under each recommendation policy. We take the average distance for each user-episode, and then average over all 100 evaluation episodes per user. Dashed lines plot the mean.

We further examine how changes in polarization across the three scenarios differ between individuals with extreme beliefs and those with more moderate initial beliefs about GMOs. We define a moderate prior to be between 0.25 and 0.75 (recall that 0.5 is neutral), and extreme otherwise.

Recall that the behavioral biases impede the ability of individuals to change their opinions about GMOs upon receiving new information. Therefore, the higher polarization relative to the scenario without these biases should have accrued because individuals with extreme beliefs failed to moderate their beliefs when presented with new information. On the other hand, with the algorithms, new information is likely to have driven the higher polarization. Consuequently, we should expect higher polarization among individuals who started with moderate beliefs as well.

We present the results from this analysis in Table 11. The results suggest that the higher polarization due to biases is mostly concentrated among individuals starting with extreme beliefs. On the other hand, both like- and time-maximizing algorithms result in higher polarization among individuals who start with moderate beliefs. Thus, biases and algorithms lead to higher polarization through different pathways.

Figure 11: Change in Polarization Among users with Moderate vs. Extreme Beliefs

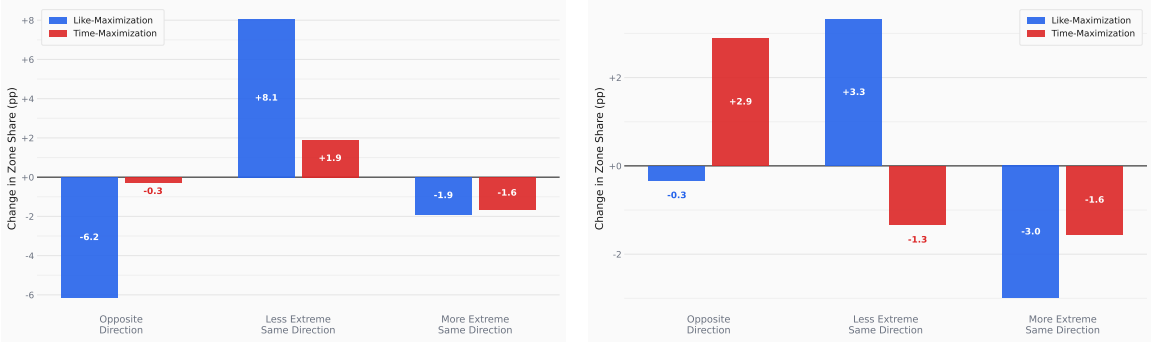


Notes. This figure plots the change in polarization among users who start with moderate vs. extreme beliefs at the beginning of experiment under different scenarios. The results suggest that biases lead to higher polarization among extreme users, while algorithms lead to higher polarization among moderates.

We also find that the time-maximizing algorithm leads to less polarization than the like-maximizing algorithm (comparing the third column to the second column in Table 8). This is despite the fact that the like-maximization achieved a higher % increase in the objective function than time-maximization

(see Figure 9). To shed light on this mechanism, we plot how the two algorithms differ in the types of content they serve to individuals with moderate versus extreme priors in Figure 12.

Figure 12: Signal Position Relative to Prior by Recommendation Policy



(a) Users with Extreme Priors (b) Users with Moderate Priors

Notes. This figure shows the change in the share of recommended signals (in percentage points) falling in each zone relative to random recommendation, in the absence of information processing biases. *Opposite Direction*: signal falls on the opposite side of 0.5 from the prior. *Less Extreme Same Direction*: signal is on the same side as the prior but closer to 0.5. *More Extreme Same Direction*: signal is on the same side as the prior but farther from 0.5.

Figure 12a shows that among users with extreme priors, like-maximization substantially increases the share of confirmatory but more moderate signals (+8.1 pp), while sharply reducing signals from the opposite direction (−6.2 pp). Signals that are more extreme in the same direction also decline slightly (−1.9 pp). This is consistent with the estimates from Table 6, as individuals are less likely to "like" articles in two zones: (a) those in the opposite direction, and (b) those in same direction as their prior but are more extreme. Because the domain involves a controversial topic where users arrive with polarized priors (see Figure 1), like-maximization serves content that reinforces the existing positions of extreme users, thereby preserving more of the initial belief polarization. This content-supply mechanism explains why the like-maximization algorithm yields higher polarization than the random recommendation. In contrast, time-maximization has much smaller effects on the signal composition for extreme users, with a modest increase in confirmatory but more moderate signals (+1.9 pp) and little change in opposite-direction signals (−0.3 pp).

Figure 12b reveals a similar pattern under like-maximization among users with moderate priors as with those with extreme priors. However, unlike extremes, under time-maximization, moderates receive a higher share of opposite-direction signals and a lower share of less-extreme same-direction signals. As a result, time-maximization exposes moderates to more disconfirmatory content. This

is consistent with the estimates of the dwell time model, which shows that individuals spend more time on disconfirmatory content relative to confirmatory, moderate content. For moderate users with priors close to 0.5, opposite-direction signals are close to the prior, so the positive main effect of opposite content on dwell time dominates the negative distance interaction, making such content engagement-maximizing.

Thus, our analysis reveals that the choice of engagement metric can shape how recommendation algorithms influence belief polarization. Like-maximization primarily preserves existing polarization by curating content closely aligned with users' prior beliefs. This induces minimal belief updating for both moderate and extreme users. Dwell-time maximization produces lower polarization than like-maximization, but through a different mechanism. For users with extreme priors, it behaves similarly to like-maximization by serving confirmatory content, while for moderates, it increases exposure to content in the opposite direction. However, this disconfirmatory content tends to be mild—close to the user's prior rather than strongly disconfirmatory. As a result, its capacity to exert a meaningful depolarizing effect is limited. Both algorithms do not outperform the random benchmark in decreasing polarization. These findings suggest that platforms face a meaningful choice in how they operationalize engagement, and that this choice has direct consequences for the distribution of beliefs in the population they serve.

7 Conclusion

Our study disentangles the roles of biases in information processing and engagement-maximizing recommendation algorithms in exacerbating the polarization of beliefs. Our main findings are the following.

First, we find that *weight distortion*—rather than *signal distortion*—is the primary channel through which information processing biases escalate polarization. Signal distortion has a negligible net effect because it symmetrically pulls perceived signals toward an individual's prior: signals more extreme than the prior are perceived as less extreme (a moderating force), while signals less extreme than the prior are perceived as more extreme (a polarizing force), and these opposing adjustments cancel out. Weight distortion, in contrast, leads individuals to systematically discount information that differs from their prior beliefs, limiting belief updating toward the center and reinforcing prior positions. This finding has significant implications for information design and policy. It suggests that polarization does not primarily arise from users misunderstanding the content they encounter,

but from how they selectively weigh information based on their prior beliefs. This challenges the sufficiency of improving perceived content accuracy (e.g., via fact-checking or credibility labels) to curb polarization. Instead, policymakers should consider how their systems shape information weighting, and not just content exposure.

Second, both like-maximizing and dwell-time-maximizing recommendation algorithms increase belief polarization relative to random content assignment, but through distinct mechanisms and to different degrees. Like-maximization serves content that is closely aligned with users' prior beliefs, particularly for users with extreme priors, where it especially reduces exposure to opposing viewpoints while increasing exposure to belief-confirming content. This anchors extreme users in their existing positions, preserving initial polarization. Dwell-time maximization, by contrast, more frequently exposes moderate users to content from the opposite side of the attitudinal spectrum. However, this cross-cutting content tends to be mild—close to the user's prior rather than strongly disconfirmatory. This limits the extent of the potential depolarizing force exerted by cross-cutting content.

Our findings have implications for how platforms monetize user attention. We show that optimizing for dwell time amplifies belief polarization less than optimizing for likes. This suggests that a shift towards dwell-time–based engagement metrics could mitigate some of the societal risks of engagement optimization, while being aligned with the platform's objective of monetizing content. More broadly, engagement strategies that emphasize reflective consumption over reactive endorsement may yield benefits for both platforms and the public. In this sense, our results highlight that the metric used to value attention is not merely a design choice. Rather, it also has potential societal externalities. Lastly, our findings are also important for the growing emphasis on data privacy regulations—including the GDPR (2018), CPRA (2020), PECR (2018), and PIPEDA (2021). While these laws rightly aim to protect users, they often restrict access to behavioral signals like dwell time—which rely on cookies—while permitting the logging of explicit but more reinforcement-prone signals such as likes, shares, etc. Our findings suggest that over-restricting passive engagement metrics could inadvertently exacerbate belief polarization by forcing platforms to rely on more polarizing forms of engagement. Policymakers may therefore wish to consider allowing the logging of dwell time through "strictly necessary" or "functional" cookies, especially when such signals help mitigate algorithmic harms.

Our study suffers from a few limitations that could offer fruitful avenues for future research. First, we have delved deep into the role of personalized engagement-optimizing recommendation

system in the polarization of beliefs for beliefs about one controversial topic - the use of GMOs in food products. Future research could consider additional controversial topics such as climate change, vaccines, abortion, gun control, etc. Interesting boundary conditions of our findings based on the nature of these topics might be awaiting discovery. For example, personalized engagement-optimizing recommendation systems might have different role for science-based topics such as climate change and vaccines compared to rights-based topics such as abortion and gun control. Perhaps belief adjustment and engagement provision behavior is different based on this dimension of content. Research along these lines could provide additional guidance to further improve the design and potential regulation of platforms' recommendation engines. Second, we have studied the role of personalized-engagement optimizing recommendation in tandem with innate human biases in information processing. Extant research has identified other innate human traits such as homophily (,i.e., the tendency to mainly interact with similar other on online platforms) as another factor that contributes to escalating belief polarization (Bakshy et al., 2015). Future research could disentangle the relative contribution of homophily from that of information gatekeeping through social network-based personalized engagement-optimizing recommendation systems in exacerbating the polarization of beliefs. Furthermore, a study that considers the joint effect of both homophily and information processing biases could further advance our understanding of the algorithmic amplification of belief polarization and provide valuable insights for more responsible design of social networking platforms.

Finally, this study abstracts away from biases in information search. Future work can extend this line of research by inferring biased search by letting users decide when they want to terminate their browsing session. Such a study would provide additional understanding of biased search and its role in accentuating the polarization of beliefs. Our paper seeks to deepen our understanding of the role of personalized engagement-optimizing algorithms in escalating belief polarization and offer guidance for more responsible platform design and potential regulation of personalized engagement-optimizing recommendation systems. We hope our findings and their implications benefit key societal stakeholders—consumers, firms, and policymakers—and stimulate further research in this critical area.

References

Ajzen, Icek, Fishbein, Martin, Lohmann, Sophie, and Albarracín, Dolores (2018). "The influence of attitudes on behavior". The handbook of attitudes, volume 1: Basic principles, 197–255.

Arora, Swapan Deep, Singh, Guninder Pal, Chakraborty, Anirban, and Maity, Moutusy (2022). "Polarization and social media: A systematic review and research agenda". Technological Forecasting and Social Change, 183, 121942.

Bakshy, Eytan, Messing, Solomon, and Adamic, Lada A (2015). "Exposure to ideologically diverse news and opinion on Facebook". Science, 348 (6239), 1130–1132.

Carother, Thomas and O'Donohue, Andrew (2019). "How to Understand the Global Spread of Political Polarization". `https://carnegieendowment.org/posts/2019/10/how-to-understand-the-global-spread-of-political-polarization?lang=en`.

Chen, Minmin, Chang, Bo, Xu, Can, and Chi, Ed H (2021). "User response models to improve a reinforce recommender system". In "Proceedings of the 14th ACM international conference on web search and data mining", 121–129.

Chen, Xiaocong, Yao, Lina, McAuley, Julian, Zhou, Guanglin, and Wang, Xianzhi (2023). "Deep reinforcement learning in recommender systems: A survey and new perspectives". Knowledge-Based Systems, 264, 110335.

Cinelli, Matteo, De Francisci Morales, Gianmarco, Galeazzi, Alessandro, Quattrociocchi, Walter, and Starnini, Michele (2021). "The echo chamber effect on social media". Proceedings of the National Academy of Sciences, 118 (9), e2023301118.

Dahlgren, Peter M (2020). "Media echo chambers: Selective exposure and confirmation bias in media use, and its consequences for political polarization".

Dandekar, Pranav, Goel, Ashish, and Lee, David T (2013). "Biased assimilation, homophily, and the dynamics of polarization". Proceedings of the National Academy of Sciences, 110 (15), 5791–5796.

Erdem, Tülin and Keane, Michael P (1996). "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets". Marketing science, 15 (1), 1–20.

Fong, Jessica, Guo, Tong, and Rao, Anita (2024). "Debunking misinformation about consumer products: Effects on beliefs and purchase behavior". Journal of Marketing Research, 61 (4), 659–681.

Freedman, Jonathan L and Sears, David O (1965). "Selective exposure". In "Advances in experimental social psychology", volume 2, 57–97. Elsevier.

Frey, Dieter (1986). "Recent research on selective exposure to information". Advances in experimental social psychology, 19, 41–80.

Guess, Andrew M, et al. (2023). "How do social media feed algorithms affect attitudes and behavior in an election campaign?" Science, 381 (6656), 398–404.

Ha, David and Schmidhuber, Jürgen (2018). "World models". arXiv preprint arXiv:1803.10122.

Hammond, Kenneth R (1948). "Measuring attitudes by error-choice: an indirect method." The Journal of Abnormal and Social Psychology, 43 (1), 38.

Huszár, Ferenc, Ktena, Sofia Ira, O'Brien, Conor, Belli, Luca, Schlaikjer, Andrew, and Hardt, Moritz (2022). "Algorithmic amplification of politics on Twitter". Proceedings of the National Academy of Sciences, 119 (1), e2025334119.

Itzchakov, Guy and Van Harreveld, Frenk (2018). "Feeling torn and fearing rue: Attitude ambivalence and anticipated regret as antecedents of biased information seeking". Journal of Experimental Social Psychology, 75, 19–26.

Iyengar, Raghuram, Ansari, Asim, and Gupta, Sunil (2007). "A model of consumer learning for service quality and usage". Journal of marketing research, 44 (4), 529–544.

Khurana, Purnima, Gupta, Bhavna, Sharma, Ravish, and Bedi, Punam (2024). "Session-aware recommender system using double deep reinforcement learning". Journal of Intelligent Information Systems, 62 (2), 403–429.

Kim, Eun-mee and Ihm, Jennifer (2020). "More than virality: Online sharing of controversial news with activated audience". Journalism & Mass Communication Quarterly, 97 (1), 118–140.

Kim, Yonghwan and Kim, Youngju (2019). "Incivility on Facebook and political polarization: The mediating role of seeking further comments and negative emotion". Computers in Human Behavior, 99, 219–227.

Krosnick, JA, Judd, CM, and Wittenbrink, Bernd (2005). "Attitude measurement". Handbook of attitudes and attitude change, 21–76.

Levy, Ro'ee (2021). "Social media, news consumption, and polarization: Evidence from a field experiment". American economic review, 111 (3), 831–870.

Mehta, Nitin, Chen, Xinlei, and Narasimhan, Om (2008). "Informing, transforming, and persuading: Disentangling the multiple effects of advertising on brand choice decisions". Marketing Science, 27 (3), 334–355.

Mnih, Volodymyr, et al. (2015). "Human-level control through deep reinforcement learning". nature, 518 (7540), 529–533.

Modgil, Sachin, Singh, Rohit Kumar, Gupta, Shivam, and Dennehy, Denis (2024). "A confirmation bias view on social media induced polarisation during Covid-19". Information Systems Frontiers, 26 (2), 417–441.

Murray, Mark and Marquez, Alexandra (2022). "Here's what's driving America's increasing political polarization". https://www.nbcnews.com/meet-the-press/meetthepressblog/s-s-driving-americas-increasing-political-polarization-rcna89559.

Narayanan, Sridhar and Manchanda, Puneet (2009). "Heterogeneous learning and the targeting of marketing communication for new products". Marketing science, 28 (3), 424–441.

Pariser, Eli (2011). The filter bubble: What the Internet is hiding from you. penguin UK.

Paulhus, DL (1991). "Measurement and control of response bias". Measures of Personality and Social Psychological Attitudes/Academic Press, Inc.

Pew Research Center (2014). "Political Polarization in the American Public". https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/.

Rabin, Matthew and Schrag, Joel L (1999). "First impressions matter: A model of confirmatory bias". The quarterly journal of economics, 114 (1), 37–82.

Rathje, Steve, Robertson, Claire, Brady, William J, and Van Bavel, Jay J (2024). "People think that social media platforms do (but should not) amplify divisive content". Perspectives on Psychological Science, 19 (5), 781–795.

Rathje, Steve, Van Bavel, Jay J, and Van Der Linden, Sander (2021). "Out-group animosity drives engagement on social media". Proceedings of the National Academy of Sciences, 118 (26), e2024292118.

Song, Zhanghan, Zhang, Dian, Shi, Xiaochuan, Li, Wei, Ma, Chao, and Wu, Libing (2021). "DEN-DQL: Quick convergent deep Q-learning with double exploration networks for news recommendation". In "2021 International Joint Conference on Neural Networks (IJCNN)", 1–8. IEEE.

Sriram, S, Chintagunta, Pradeep K, and Manchanda, Puneet (2015). "Service quality variability and termination behavior". Management Science, 61 (11), 2739–2759.

Sunstein, Cass R (2009). Going to extremes: How like minds unite and divide. Oxford University Press.

Törnberg, Petter (2022). "How digital media drive affective polarization through partisan sorting". Proceedings of the National Academy of Sciences, 119 (42), e2207159119.

Van Hasselt, Hado, Guez, Arthur, and Silver, David (2016). "Deep reinforcement learning with double q-learning". In "Proceedings of the AAAI conference on artificial intelligence", volume 30.

Wang, You-Qiang and Tsui, Kai-Yuen (2000). "Polarization orderings and new classes of polarization indices". Journal of Public Economic Theory, 2 (3), 349–363.

Wolfson, Michael C (1994). "When inequalities diverge". The American Economic Review, 84 (2), 353–358.

Wolfson, Michael C (1997). "Divergent inequalities: theory and empirical results". Review of Income and Wealth, 43 (4), 401–421.

Xia, Xue, Gu, Neng, Badani, Dhruvil Deven, and Zhai, Andrew (2022). "How Pinterest Leverages Realtime User Actions in Recommendation to Boost Homefeed Engagement Volume". Pinterest Engineering Blog.

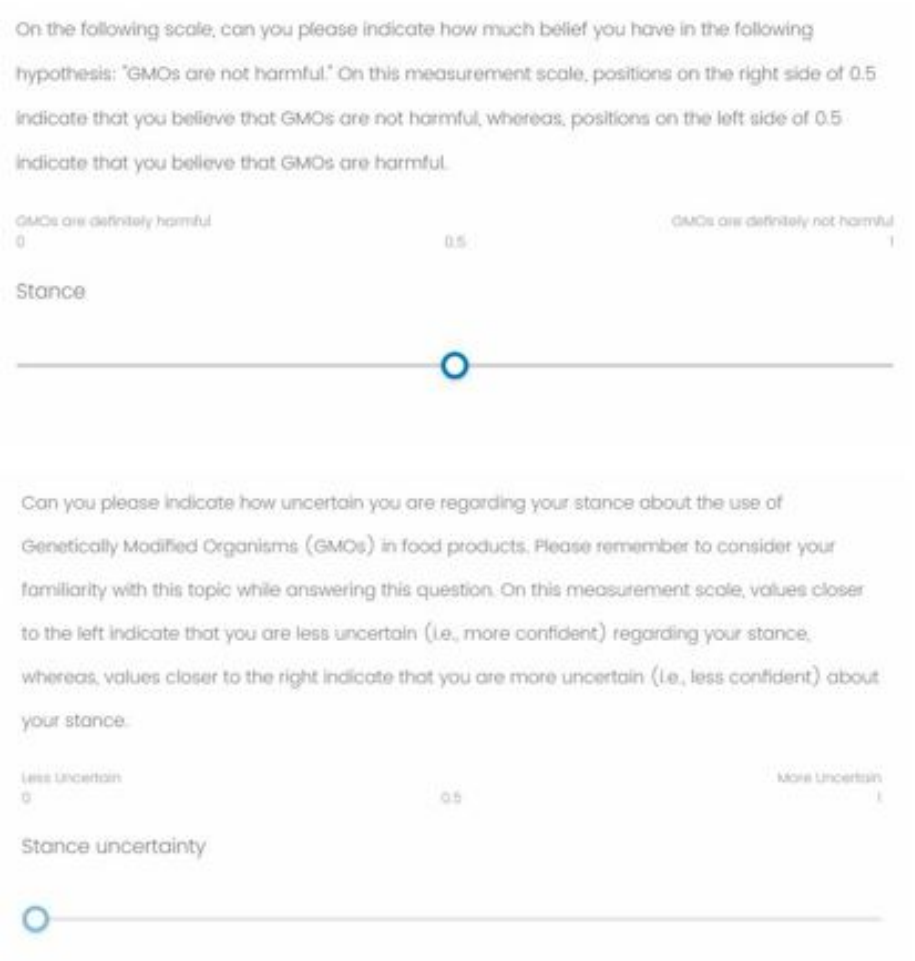
Yarchi, Moran, Baden, Christian, and Kligler-Vilenchik, Neta (2021). "Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media". Political Communication, 38 (1-2), 98–139.

APPENDIX

A Experimental Design and Sample

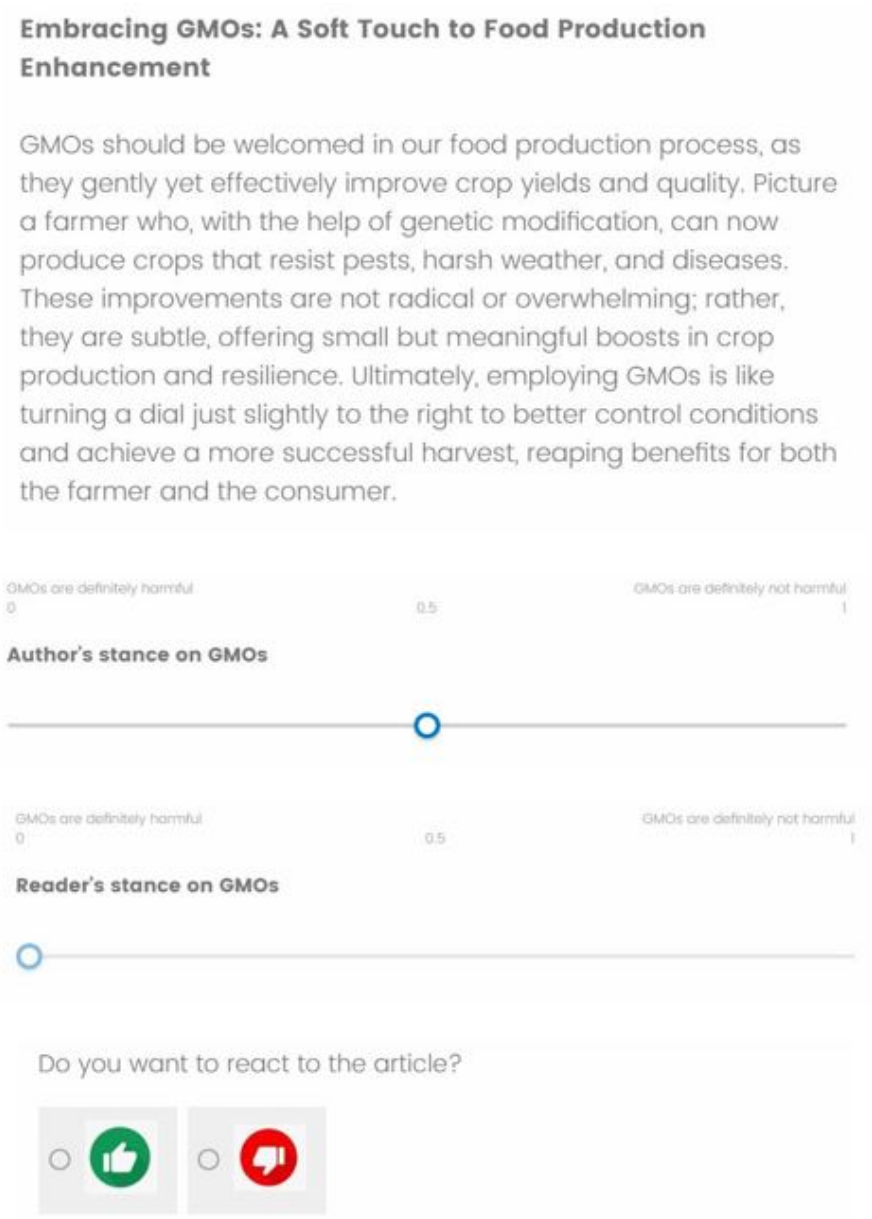
A.1 Prior Elicitation

Figure A.1: Elicitation of prior beliefs



Notes. Prior to the experimental content browsing session, we elicit their prior belief about the harmfulness of GMOs (*PriorMean*) as well as the uncertainty they have in their prior belief (*PriorVariance*).

Figure A.2: Example of a period in a user's content browsing session.



Notes. First, we measure an individual's perception of the belief reflected in the recommended article. Next, we measure the individual's current belief in the hypothesis that "GMOs are not harmful." Finally, we log the individual's explicit feedback provision decision (i.e., their decision to like, dislike, or not react to the recommended article.). Note that we also log the individual's implicit feedback provision decision (i.e., time spent by the individual while reading the article).

A.2 Sample Demographics

In this appendix, we present the distribution of a few key demographic traits of the participants recruited for our study. This data was collected in the first wave of our study and then used for the second wave of our study. We recruited a representative sample of US participants from Prolific.com.

Figure A.3: Distribution of Income

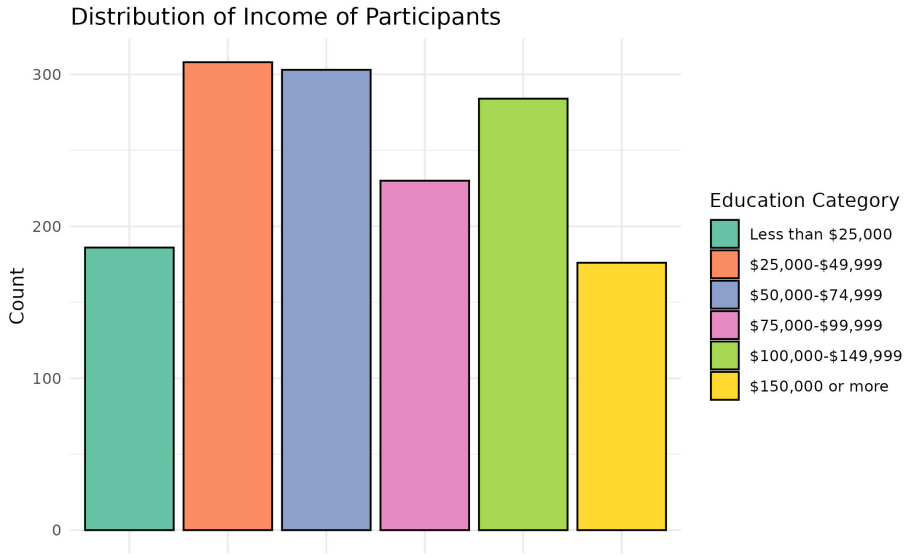


Figure A.4: Distribution of Education

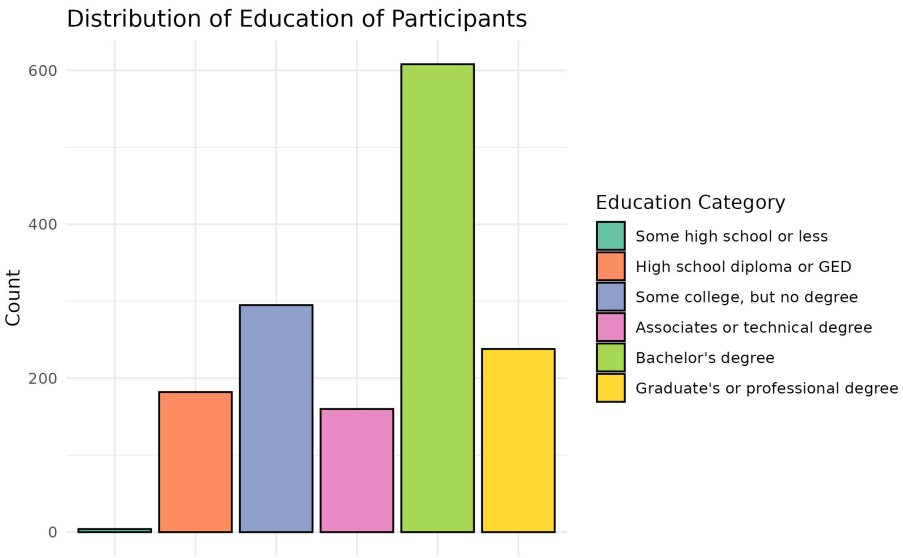
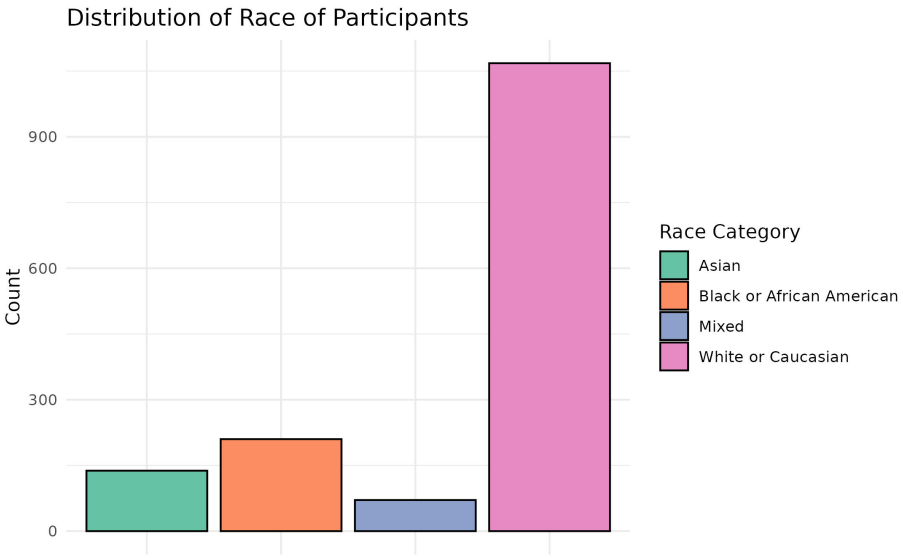


Figure A.5: Distribution of Race



B Supplementary Analysis

B.1 Signal Misperceptions

We find that individuals also misinterpret signals in a systematic way that is not related to their prior belief. Figure 3 plots the average perceived signal by the true signal. Recall that true signals of greater than 0.5 indicate the individual is pro-GMO while values of less than 0.5 indicate anti-GMO. Figure 3 shows individuals tend to perceive content about GMOs as more pro-GMO that it actually is, as average perceived signal is usually higher than the average true signal (represented by the line). The misperception is greater for pro-GMO content; individuals perceive this content to be more extreme than it is.

While this demonstrates that individuals indeed misinterpret the signals in a systematic way, this is not the kind of signal distortion we focus on in the paper, as this kind of misperception likely varies across contexts (e.g., topic domain), and therefore is less likely to generalize beyond our experimental setting. Our main focus is how signal distortion varies by the prior belief because this is more relevant to information processing biases such as confirmation bias. Therefore, when we measure signal distortion, we difference out this kind of systematic misinterpretation by including signal bin fixed effects (e.g., Equation 3).

B.2 Reading Time and Article Length

Table B.1: Reading Time and Article Length

	Log(Dwell Time) (1)
Constant	-1.707***
	(0.3301)
log(N Words)	0.7215***
	(0.0490)
Observations	14,880
R^2	0.01400
Dependent variable mean	3.1737

Notes. This table reports OLS estimates of the relationship between reading time (dwell time) and article length. The dependent variable is the natural log of reading time in seconds. The independent variable is the log of the number of characters in the article. Standard errors, reported in parentheses, are clustered at the participant level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

C Model Derivations

C.1 Dwell Time

Recall that in Equation 9, we specified the individual's dwell time decision problem as follows:

$$\tau_{it} \in \arg\max_{\tau_{it}>0} \left\{ a(x_{it})\, \eta_{it}\, \log(1+\tau_{it}) - \exp(c_i + f(N_{it}))\, \tau_{it} \right\}.$$

Since the individual is maximizing the objective function specified above using the decision variable τ_{it}, the first-order condition for an interior optimum ($\tau > 0$) is:

$$\frac{\partial U(\tau_{it})}{\partial \tau_{it}} = \frac{a(x_{it})\, \eta_{it}}{1+\tau_{it}} - \exp(c_i + f(N_{it})) = 0. \tag{15}$$

Solving (15) yields

$$1 + \tau_{it} = \frac{a(x_{it})\, \eta_{it}}{\exp(c_i + f(N_{it}))}. \tag{16}$$

Taking logs gives

$$\log(1+\tau_{it}) = \log a(x_{it}) - (c_i + f(N_{it})) + \epsilon_{it}, \tag{17}$$

where $\epsilon_{it} \equiv \log(\eta_{it})$. Substituting $\log a(x_{it}) = \delta_0 + \Delta x_{it}$ from (6), we obtain the estimating equation:

$$\log(1+\tau_{it}) = \delta_0 + \Delta x_{it} - c_i - f(N_{it}) + \epsilon_{it}. \tag{18}$$

This expression corresponds directly to estimation Equation 10 in the main text.

C.2 Explicit Feedback

Recall that we let $y_{it} \in \{-1, 0, 1\}$ denote the observed reaction, where $y_{it} = -1$ corresponds to "dislike", $y_{it} = 0$ to *no reaction*, and $y_{it} = 1$ to "like". In addition, we model choices via two thresholds $\delta_d < \delta_l$ such that:

$$y_{it} = \begin{cases} -1 & \text{if } u_{it} < \delta_d, \\ 0 & \text{if } \delta_d \le u_{it} \le \delta_l, \\ 1 & \text{if } u_{it} > \delta_l. \end{cases} \tag{19}$$

Further, we assumed that ξ_{it} follows a logistic distribution with CDF as follows:

$$\Lambda(v) \;\equiv\; \Pr(\xi_{it} \le v) = \frac{1}{1 + e^{-v}}. \tag{20}$$

In this appendix, we derive the estimation equation resulting from our model. Using the latent-index representation $u_{it} = \beta' x_{it}^F + \xi_{it}$ and the threshold rule (19), the conditional choice probabilities (conditional on x_{it}^F) are obtained by rewriting each event in terms of the shock ξ_{it} and applying the logistic CDF $\Lambda(\cdot)$:

$$\Pr(y_{it} = -1 \mid x_{it}^F) = \Pr\big(u_{it} < \delta_d \mid x_{it}^F\big) = \Pr\big(\beta' x_{it}^F + \xi_{it} < \delta_d \mid x_{it}^F\big) = \Pr\big(\xi_{it} < \delta_d - \beta' x_{it}^F \mid x_{it}^F\big) = \Lambda(\delta_d - \beta' x_{it}^F), \tag{21}$$

$$\begin{aligned} \Pr(y_{it} = 0 \mid x_{it}^F) &= \Pr\big(\delta_d \le u_{it} \le \delta_l \mid x_{it}^F\big) = \Pr\big(\delta_d - \beta' x_{it}^F \le \xi_{it} \le \delta_l - \beta' x_{it}^F \mid x_{it}^F\big) \\ &= \Lambda(\delta_l - \beta' x_{it}^F) - \Lambda(\delta_d - \beta' x_{it}^F), \end{aligned} \tag{22}$$

$$\begin{aligned} \Pr(y_{it} = 1 \mid x_{it}^F) &= \Pr\big(u_{it} > \delta_l \mid x_{it}^F\big) = \Pr\big(\beta' x_{it}^F + \xi_{it} > \delta_l \mid x_{it}^F\big) = \Pr\big(\xi_{it} > \delta_l - \beta' x_{it}^F \mid x_{it}^F\big) \\ &= 1 - \Pr\big(\xi_{it} \le \delta_l - \beta' x_{it}^F \mid x_{it}^F\big) = 1 - \Lambda(\delta_l - \beta' x_{it}^F). \end{aligned} \tag{23}$$

Equations (21)–(23) define an ordered logit model with cutpoints $\delta_d < \delta_l$ and index $\beta' x_{it}^F$.

For $k \in \{-1, 0\}$, the cumulative probability implied by the same threshold structure is

$$\Pr(y_{it} \le k \mid x_{it}^F) = \begin{cases} \Pr(y_{it} = -1 \mid x_{it}^F), & k = -1, \\ \Pr(y_{it} \in \{-1, 0\} \mid x_{it}^F), & k = 0, \end{cases}$$

$$
\begin{aligned}
&= \begin{cases} \Pr(u_{it} < \delta_d \mid x_{it}^F), & k = -1, \\ \Pr(u_{it} \le \delta_l \mid x_{it}^F), & k = 0, \end{cases} \\
&= \begin{cases} \Pr(\beta' x_{it}^F + \xi_{it} < \delta_d \mid x_{it}^F), & k = -1, \\ \Pr(\beta' x_{it}^F + \xi_{it} \le \delta_l \mid x_{it}^F), & k = 0, \end{cases} \\
&= \begin{cases} \Pr(\xi_{it} < \delta_d - \beta' x_{it}^F \mid x_{it}^F), & k = -1, \\ \Pr(\xi_{it} \le \delta_l - \beta' x_{it}^F \mid x_{it}^F), & k = 0, \end{cases} \\
&= \begin{cases} \Lambda(\delta_d - \beta' x_{it}^F), & k = -1, \\ \Lambda(\delta_l - \beta' x_{it}^F), & k = 0. \end{cases}
\end{aligned} \tag{24}
$$

Equivalently, for each cutoff $k \in \{-1, 0\}$,

$$
\Pr(y_{it} \le k \mid x_{it}^F) = \Lambda(\delta_k - \beta' x_{it}^F), \qquad \text{where } \delta_{-1} \equiv \delta_d,\ \delta_0 \equiv \delta_l. \tag{25}
$$

Recall that $\Lambda(v) = \frac{1}{1+e^{-v}}$ is the logistic CDF. Let:

$$
p_{it}(k) \ \equiv \ \Pr(y_{it} \le k \mid x_{it}^F).
$$

Then (25) implies:

$$
p_{it}(k) = \frac{1}{1 + \exp\big(-(\delta_k - \beta' x_{it}^F)\big)}. \tag{26}
$$

Taking reciprocals and subtracting 1 on both sides of (26) gives:

$$
\frac{1}{p_{it}(k)} - 1 = \exp\big(-(\delta_k - \beta' x_{it}^F)\big). \tag{27}
$$

Inverting both sides:

$$
\frac{p_{it}(k)}{1 - p_{it}(k)} = \exp\big(\delta_k - \beta' x_{it}^F\big). \tag{28}
$$

Finally, taking logs of both sides of (28) yields the ordered logit regression:

$$
\log\left(\frac{\Pr(y_{it} \le k \mid x_{it}^F)}{1 - \Pr(y_{it} \le k \mid x_{it}^F)}\right) = \delta_k - \beta' x_{it}^F, \qquad k \in \{-1, 0\}. \tag{29}
$$

Equation (29) is the ordered logit model: the slope on x_{it}^F is common across cutpoints (the proportional-

odds restriction), while the intercept varies across cutoffs via δ_k.